Instituto Superior de
Engenharia de Coimbra

Departamento
de Engenharia Informática e Sistemas

Classificação automática de textos através reconhecimento de entidades

Relatório de estágio apresentado para a obtenção do grau de Mestre em Informática e Sistemas

Autor
Paulo Alexandre Fernandes Costa

Orientadores
Prof. Doutor Carlos Manuel Jorge da Silva Pereira
ISEC
Eng.º Pedro Miguel Henriques dos Santos Pimentel Ladeira
CISION Portugal

Coimbra, dezembro 2014

AGRADECIMENTO

Todo o estudo e investigação que desenvolvi é sem dúvida fruto não só do meu empenho e afinco individual mas também do esforço de muitas outras pessoas que tornaram possível a conclusão de mais esta etapa tão importante quer na minha vida profissional como pessoal.

Desde já manifesto o meu agradecimento sincero a todos aqueles que me encorajaram e tornaram possível a conclusão desta etapa.

Quero agradecer ao Professor Doutor Carlos Pereira pelo conhecimento que me transmitiu, pela dedicação e o apoio prestado no decorrer de todo o período de estágio.

Agradeço a todos os meus colegas do departamento do IT da CISION Portugal que de uma forma ou de outra me ajudaram a desenvolver o meu estágio, mas dirijo um especial agradecimento: ao Eng.º António Silva, à minha equipa (Ana Duarte, Andreia Dias e António Júlio), Dr.ª Telma Serra, Eng.º Diogo Rodrigues e à Dr.ª Marllene Silva.

Um agradecimento ao Eng.º Pedro Ladeira pelo apoio, dedicação e conhecimento transmitido durante o meu estágio, à Dr.ª Ana Cláudia Duarte pela sua disponibilidade na revisão a nível linguístico desta tese, Eng.º João Ramos pela disponibilidade e ao António Pessoa pela ajuda no desenvolvimento gráfico dos diagramas.

Por fim, os últimos mas os mais importantes: a minha família, em especial a minha esposa Filipa pela ajuda e apoio incondicional que demonstrou e por me proporcionar todas as condições necessárias para que eu conseguisse atingir o meu objetivo! Para terminar quero deixar um grande beijo ao João Afonso e outro para o Miguel Pedro, os meus dois lindos filhos!

RESUMO

A sociedade de informação é um conceito que surge no fim do Século XX e que está diretamente relacionado com a Globalização. Neste contexto podemos considerar que a sociedade se encontra num processo contínuo de formação e expansão.

Com o surgimento das redes sociais e com a evolução tecnológica de vários tipos de dispositivos inteligentes assistiu-se a um crescimento alucinante na partilha de um conjunto vasto de informação. Assim, a gestão desta torna-se difícil ou praticamente impossível, sem utilização de ferramentas que permitam filtrar o que realmente é importante para o contexto das organizações, para que estas possam identificar novas oportunidades de negócio.

Neste contexto sendo a CISION Portugal líder de mercado na monitorização e segmentação de informação, urge a necessidade de melhorar o seu processo de produção, para uma melhor resposta às exigências do mercado.

Esta investigação apresenta então um novo método de indexação e segmentação de conteúdos para ser aplicado no processo interno de produção da CISION Portugal, tendo por base a identificação automática de entidades nos textos jornalísticos produzidos online e a sua relevância para um determinado tema.

Verificou-se que o novo método para temas com pouca ambiguidade funciona e consegue resultados semelhantes aos que atualmente existem, sendo que as propostas de indexação chegam mesmo a atingir um grau de certeza próximo dos 100%.

Para os temas com um elevado grau de ambiguidade e que, por sua vez, exigem uma equipa de validação de conteúdos, a investigação abordou a questão para um tema específico, através do desenvolvimento de um sistema de classificação automática de texto, utilizando para tal algoritmos probabilísticos.

A seleção do conjunto treino, para os sistemas anteriormente referidos, foi criado sem recorrer ao histórico interno produzido pelas equipas da CISION Portugal, utilizando apenas a identificação das entidades. Os resultados obtidos quando comparados com os atuais demonstram que é possível reduzir o número de propostas irrelevantes e fazer a indexação de conteúdos sem necessidade de recorrer a uma supervisão inicial por parte de uma equipa.

Palavras Chave: Classificação de texto, DbPedia e NER

ABSTRACT

The concept of information society appeared at the end of the twentieth century and is directly related to globalization. In this context, we can consider that society is in a continuous process of development and growth.

With the rise of social networks and technological advances of several types of intelligent devices there has been an incredible increase in sharing a wide range of information. Thus, managing shared information becomes difficult or practically impossible without using tools that enable to filter what is really important for the context of organizations, so that they can identify new business opportunities.

In this context, and since CISION Portugal is market leader in media monitoring and segmentation of information, there is the need to improve their production process to better respond to market demands.

This research therefore presents a new method of content tagging and segmentation to be applied to CISION Portugal's internal production process, based on the automatic identification of entities in online newspaper articles and their relevance to a particular topic.

Research showed that the new method for topics with little ambiguity works and reaches similar results to those that currently exist, with indexing proposals that reach levels of certainty of approximately 100%.

For subjects with a high degree of ambiguity which, in turn, call for a content validation team, research addressed the issue for a specific topic by developing an automatic text classification system using probabilistic algorithms.

Selection of the training group for the above mentioned systems was created without resorting to the internal historical database produced by CISION Portugal teams and only using the identification of entities. When compared with current results, the results obtained with this research show that it is possible to reduce the number of irrelevant proposals and to index content without needing initial supervision by a team.

Keywords: Text Classification, DbPedia e NER

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

ÍNDICE DE GRÁFICOS

ABREVIATURAS

WISE	–	Web Intelligence Search Engine
XML	–	E**X**tensible **M**arkup **L**anguage
DS	–	**D**bpedia **S**potlight
SOLR	–	Apache Solr
REST	–	**Re**presentational **S**tate **T**ransfer
HTTP	–	**H**ypertex**t** **T**ransfer **P**rotocol
POST	–	HTTP POST
API	–	**A**pplication **P**rogramming **I**nterface
NLP	–	**N**atural **L**anguage **P**rocessing
KDD	–	**K**nowledge **D**iscovery in **D**atabases
OCR	–	**O**ptical **C**haracter **R**ecognition
EMAA	–	**E**uropean **M**edia **A**nalysts **A**ssociation
FIBEP	–	**F**édération **I**nternationale des **B**ureauxs d'**E**xtraits de **P**resse
AMEC	–	**A**ssociation for **M**easurement and **E**valuation of **C**ommunication
SIIA	–	**S**oftware & **I**nformation **I**ndustry **A**ssociation (Associação da Indústria de Software e Informação)
DVD	–	**D**igital **V**ersatile **D**isc
AD	–	**Á**rvores de **D**ecisão
NLTK	–	Natural Language Toolkit

1. INTRODUÇÃO

Nos últimos anos, o volume de informação produzida pelos meios de comunicação tem vindo a aumentar consideravelmente, em especial o conteúdo publicado na Internet, sendo esta considerada atualmente o segundo meio de comunicação mais importante na sociedade.

Nos últimos 25 anos houve uma palavra que ecoou nos nossos ouvidos: Globalização. Muitos analisam este movimento como sendo o resultado da pós-Segunda Guerra Mundial ou como o resultado da revolução tecnológica que ocorreu nas comunicações no fim do Século XX. Mas, ao contrário do que se pensa, este processo teve início no Século XV com a descoberta do caminho marítimo para a Índia pelos portugueses (Rodrigues, et al., 2007) e foi impulsionado por acontecimentos marcantes na nossa História, tais como, a Revolução Industrial, as duas Grandes Guerras Mundiais, a queda do Muro de Berlim e a revolução tecnológica, que favoreceram o desenvolvimento da sociedade no último século, permitindo a massificação dos computadores, dos telemóveis e da Internet.

A Globalização, tal como a conhecemos, só foi possível devido à comunicação, sendo esta o elo que permite a interação entre pessoas, empresas e países. Com a evolução das comunicações através da massificação da internet, dos telemóveis e com a rapidez de troca de informação entre os vários atores, os conceitos como "Aldeia Global" e "Sociedade Conhecimento" surgiram no nosso dia-a-dia, estando intrinsecamente associados a expressões como volume de informação, rapidez na partilha e acesso à informação.

Para termos uma ideia do volume de informação que está atualmente a ser gerada e consumida, tomemos em consideração a evolução do tráfego global da internet (em petabytes por mês) versus utilizadores da internet:

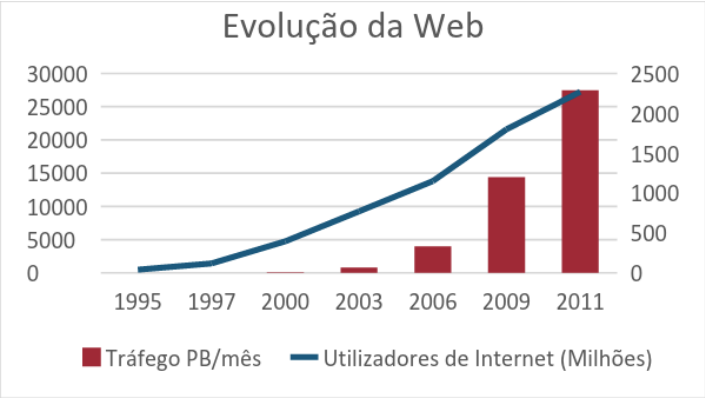


Gráfico 1.1– Evolução da Web: tráfego vs utilizadores (Vizzuality, 2011)

De acordo com o Gráfico 1.1 (Vizzuality, 2011), no ano de 1997 o tráfego gerado era de 5.4 PB\mês, para armazenar toda esta informação em DVD's seriam necessários 1,2 milhões. Dez anos mais tarde, para armazenar o tráfego gerado em DVD's e caso estes fossem organizados lado a lado, a fila teria 18,5 vezes a distância que o caminho-de-ferro Transiberiano tem. Por volta de 2009, o valor a armazenar já seria 14414 PB\mês o que significaria ter a Lua ligada à Terra por uma pilha de DVD's.

Com esta dinâmica, uma organização para estar atualizada, necessita de ter ferramentas que lhe façam a triagem e a extração da informação relevante, para que esta continue a ser competitiva no mercado global.

No passado, uma organização apenas tinha necessidade de monitorizar os seus concorrentes, os mercados financeiros, a sua rede de agentes e em alguns casos a opinião pública, por forma ajustar e delinear a sua estratégia. Mais uma vez a Internet veio alterar os hábitos instalados, com a proliferação das redes sociais, onde os consumidores comuns partilham a sua opinião sobre os produtos, as marcas e as organizações, influenciando os outros consumidores nas suas decisões. Torna-se portanto imperativo que as organizações adicionem à sua rede de monitorização, a Internet.

Segundo a consultora Nielsen, empresa de estudos de mercado (Nielsen, 2012), durante o ano de 2012, os consumidores norte-americanos navegaram na Internet em média 3 horas e 6 minutos por dia, quando no ano de 2010, esse valor se situava em 2 horas e 34 minutos. A Internet estabeleceu-se como o segundo meio de comunicação mais importante na sociedade, atrás da televisão, que continua a ser a fonte número um, onde em média cada norte-americano consome 5 horas diárias.

Como o gráfico seguinte demonstra, os consumidores gastaram em média 37 minutos em serviços, tais como, o Facebook, Twitter ou o Linkedin. Outro dado curioso é o tempo gasto pelos utilizadores em visualização de vídeos em serviços como o Youtube, Netflix ou Hulu, chegando mesmo a ultrapassar o tempo gasto em pesquisas.

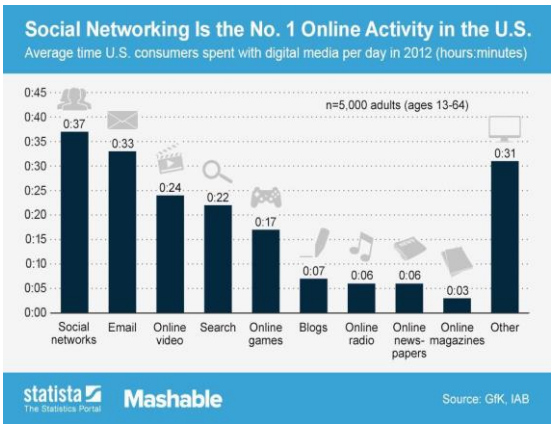


Figura 1.1 – State of the Media: The Social Media Report 2012

Mais alguns factos sobre as redes sociais (Fonseca, 2013):

- Segundo a SocialBakers, o Facebook tem mais de mil milhões de utilizadores. Em Portugal conta com 4,7 milhões de utilizadores e é o 39º país com mais utilizadores nesta rede social, já considerada a maior rede do mundo;
- Outro estudo feito pela Emarketer, demonstra que mais metade do conteúdo partilhado é via Facebook;
- O Twitter partilhou que tem mais de 200 milhões de utilizadores ativos e são publicados na sua rede mais de 500 milhões de tweets diários;
- O Youtube tem mais de mil milhões de utilizadores ativos, por minuto são publicadas 48 h de vídeo;
- A Cisco publicou que 90% dos jovens admitem que a primeira coisa que fazem quando acordam é ver o email, as mensagens e as atualizações das redes sociais via smartphones.

Com esta quantidade de informação publicada e partilhada na Internet, onde tudo acontece a uma velocidade estonteante, associada à mudança de hábitos por parte dos consumidores, as organizações sentiram-se pressionadas e ao mesmo tempo motivadas a modificar a sua maneira de comunicar e a forma de monitorizar o meio envolvente, na tentativa de ganhar vantagem relativamente aos seus concorrentes. Como consequência destas mudanças, atualmente as organizações enfrentam alguns desafios:

Será esta quantidade de informação realmente relevante para uma organização por forma dar-lhe vantagem e competitividade relativamente aos seus concorrentes?

- Como gerir grandes volumes de informação?
- Como aceder em tempo útil?

Neste contexto, temos que considerar o papel preponderante que a CISION desempenha sendo considerada líder mundial nas áreas da monitorização e avaliação de informação veiculada nos meios de comunicação social. Com escritórios na Europa, América e Ásia, disponibiliza serviços integrados de Media Intelligence e Reputation Management em 90 países em simultâneo.

Estando a CISION Portugal inserida neste grupo e a competir num mercado global, é de certa forma forçada a apostar na investigação, com o objetivo de melhorar\otimizar os seus processos de monitorização e conseguir novas formas de extrair informação das suas bases de dados e torná-la acessível aos seus clientes.

Surge desta forma a necessidade de desenvolver um sistema que permitirá a identificação das Entidades referidas no conteúdo produzido pelos meios de comunicação.

Por Entidades deveremos entender referências a Personalidades, Marcas e Empresas, cujos nomes são muitas vezes palavras comuns ou poderão aparecer em contextos não relacionados com as referidas Entidades.

O conhecimento sobre cada uma delas deverá ser obtido a partir de definições disponíveis em recursos online (exemplo: www.dbpedia.org).

1.1. Apresentação da Empresa

A origem do grupo CISION remonta a 1892, quando era um departamento de publicidade da Svenska Telegrambyrån, uma empresa sueca de serviços de Clipping. Ao longo dos anos, expandiu a sua atividade pela Europa e América do Norte. Antes de 2007, o grupo operava um pouco por todo o mundo sob diferentes nomes (como Observer, Romeike, Bacon's, Bowdens e Memorandum), o que levou a iniciar um processo de rebranding, passando a chamar-se CISION.

Hoje, a CISION é líder mundial na disponibilização de serviços para planeamento, contacto, monitorização e análise de media. Com escritórios na Europa, América e Ásia disponibiliza serviços integrados de Media Intelligence e Reputation Management em 90 países em simultâneo. Conta atualmente com 50.000 clientes, integra 2.700 profissionais e monitoriza mais de 200.000 fontes de informação em todo o mundo. A CISION está cotada na Bolsa de Valores de Estocolmo e tem aproximadamente 20.000 acionistas.

Desde de 2007, as plataformas da CISION já foram distinguidas por cinco vezes com os prémios da SIIA (Associação da Indústria de Software e Informação). Os prémios CODiE conquistados foram:

- 2013 - Best Media and Information Monitoring Solution;
- 2012 - Best Online Business Information Service;
- 2011 - Best Marketing/PR Solution;
- 2010 - Best Social Media Aggregation Service;
- 2009 - Best Online News Service.

A CISION é membro da FIBEP (Federation Internationale des Bureaux D'extraits de Presse), da IABM (International Association of Broadcast Monitors), da AMEC (Association for Measurement and Evaluation of Communication) e Secretário Geral da EMAA (European Media Analysts Association).

A CISION é em Portugal e na Península Ibérica uma referência na área de análise de meios de comunicação social e tem a responsabilidade de seleção, tratamento e análise de informação

em Portugal, Espanha e América Latina para o grupo CISION e diretamente para várias empresas internacionais.

Todo o sistema de produção é assistido por computador. Toda a informação é digitalizada e tratada em suporte digital, recorrendo a métodos eletrónicos de pesquisa em texto (OCR), decomposição de imagem e reconhecimento de voz, validados por técnicos documentalistas. Numa primeira fase o controlo de qualidade é assegurado pelo documentalista e numa segunda fase pelo gestor de cliente.

Todos os analistas da CISION Portugal têm formação académica superior nas áreas da Comunicação, Jornalismo e Marketing e formação específica ministrada pelo Grupo CISION de forma a se qualificarem para a realização destes trabalhos.

Os trabalhos de análise obedecem às metodologias desenvolvidas pelo Grupo CISION reconhecidas internacionalmente.

Todos os processos de trabalho são sujeitos a controlo de qualidade e verificação interna.

1.1.1. Monitorização

Monitorização, segmentação e classificação da informação veiculada na imprensa, televisão, rádio e Internet sobre o cliente, e/ou assuntos relativos. A CISION Portugal integra todos os serviços de numa plataforma única e disponível na Internet.

Os meios de comunicação monitorizados pela CISION Portugal encontram-se referenciados no Quadro 1.1.

Quadro 1.1 - Meios analisados pela CISION Portugal

Tipo de Meio	Meios
Imprensa	Nacional, Revistas, Especializada, Regional, Regiões Autónomas e principal imprensa estrangeira. Cerca de 1.200 meios de informação escrita
Televisão	AR TV, Benfica TV, A Bola TV, CM TV, Económico TV, PORTO Canal, RTP1, RTP2, RTP Informação, RTP Açores, RTP África, RTP Madeira, SIC, SIC Mulher, SIC Notícias, SIC Radical, SportTV 1, SportTV 2, SportTV 3, SportTV Live, TVI, TVI 24
Rádio	Antena 1, Rádio Comercial, Renascença, TSF
Internet	Cerca de 90000 meios de informação online com redação própria de conteúdos
Imprensa Africana	Principais meios comunicação de Angola e Moçambique

1.2. Objetivos

A investigação teve como objeto de estudo a viabilidade da implementação de um sistema automático de indexação e segmentação de textos jornalísticos através da identificação automática de Entidades (Personalidades, Marcas, Empresas\Organizações e conceitos)

O sistema deverá identificar uma Entidade e associar-lhe um peso de forma a podermos verificar a sua relevância no âmbito do texto. Estes deverão estar em língua portuguesa e serem artigos do tipo internet, uma vez, que neste meio os textos estão mais completos e corretos.

O sistema atual de produção faz indexação dos artigos por áreas, que são definidas pela combinação booleana de palavras-chave. Existindo, por isso, uma necessidade constante de afinação destas e de recorrer a uma equipa de validação para aferir a qualidade dos resultados propostos.

Pretende-se desenvolver um novo processo de indexação sem que seja utilizado o histórico de artigos já validados pelas equipas de produção para as áreas, de forma a permitir avaliar a possibilidade de criar novas áreas sem haver a necessidade inicial de serem supervisionadas pelas equipas de validação.

O estágio dividiu-se em três fases:

- Análise do atual sistema, ao levantamento de requisitos e quais os objetivos que se pretendiam;
- Levantamento de sistemas\algoritmos que pudessem ser aplicados no estudo;
- Desenvolvimento do Protótipo:
  - Implementação;
  - Validação;
  - Conclusões.

1.3. Estrutura do Relatório de Estágio

No capítulo 1 são descritos os problemas e os desafios com que a CISION Portugal atualmente se depara, por forma a enquadrar a investigação e os seus objetivos, bem como a motivação da escolha deste assunto a investigar.

Durante o capítulo seguinte é feita a apresentação do problema num contexto global da organização.

No capítulo 3 expõe-se as técnicas, as metodologias utilizadas e referenciadas no desenvolvimento da solução apresentada no capítulo 4.

A evolução da investigação é demonstrada no capítulo 4, bem como os respetivos resultados finais.

Por fim são apresentadas as conclusões obtidas durante a investigação, assim como as recomendações e linhas orientadoras para a desenvolver futuramente.

2. Apresentação do problema

Neste capítulo é feita a apresentação do problema num contexto global da organização.

2.1. Sistema de Produção

A área da produção da CISION para a segmentação de informação divide-se em três departamentos (Figura 2.1):

- Imprensa;
- Internet;
- Rádio;
- Televisão.

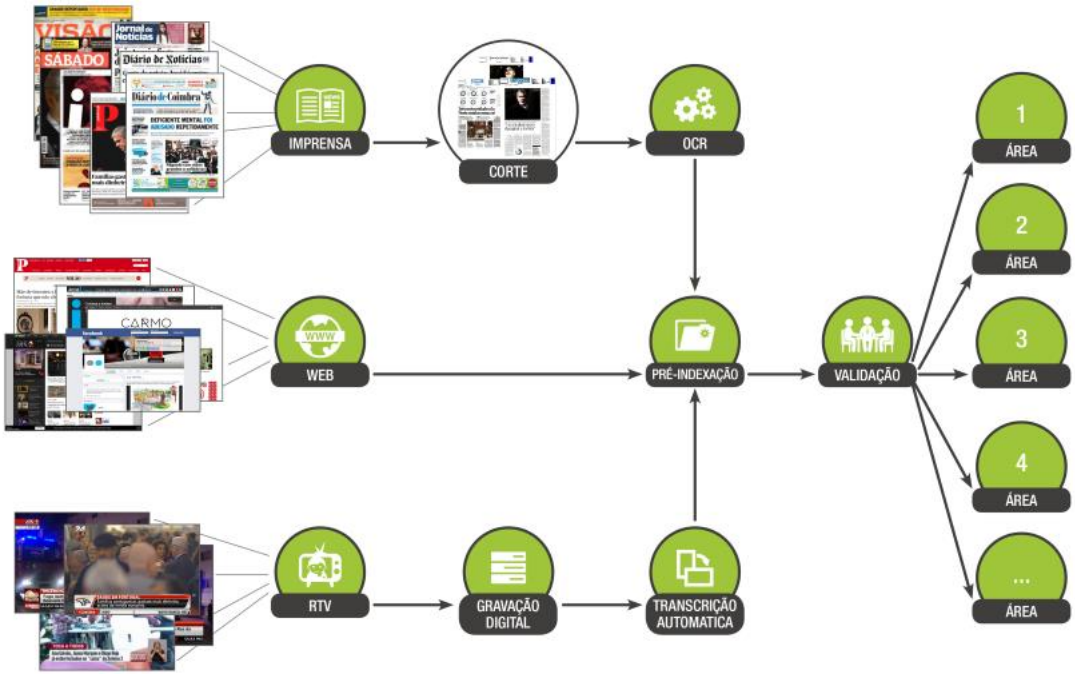


Figura 2.1 – Esquema da produção e segmentação de conteúdos

O departamento de imprensa subdivide-se em três equipas: os digitalizadores, os cortadores e os validadores. Os primeiros digitalizam página a página os jornais. Os cortadores "cortam" o jornal em notícias e por fim, os validadores validam as áreas que foram indexadas pelo motor de indexação automática.

A primeira parte do processo de segmentação começa pela digitalização dos conteúdos, mas esta divide-se ainda em dois tipos:

1. Jornais em formato digital (PDF):
    - Transformação em imagem;
    - Processo de corte notícia a notícia;
    - Extração da notícia e do texto do ficheiro inicial;
    - Geração da notícia em formato PDF.
2. Jornais em formato papel:
    - Digitalização do integral do jornal;
    - Processo de corte notícia a notícia;
    - Extração do texto através de OCR;
    - Geração do ficheiro em formato PDF.

Após a extração do texto e respetiva indexação numa base dados, o processo é igual para ambos os tipos. Os motores procedem à indexação das notícias às áreas com base em palavras-chave combinadas entre elas por operadores lógicos. O resultado desta operação é posteriormente validado pela equipa de validadores. Finalizado este processo, as notícias estão prontas a serem disponibilizadas num portal personalizado de acordo com o cliente. Em média este departamento é responsável pela organização de 4000 notícias diárias.

Para a internet, a empresa desenvolveu um sistema de monitorização automática de sítios, Web Intelligence Search Engine (WISE), que produzem conteúdos próprios. O WISE monitoriza diariamente 90000 sítios a nível global, de onde são extraídos em média 2 milhões de notícias por dia. Para que uma notícia seja considerada válida, o sistema tem que identificar o título, o corpo da notícia, a data de publicação e o sítio. Todo o conteúdo recolhido pelo WISE fica disponível para a indexação a clientes. Diariamente, apenas 10500 notícias de internet chegam à equipa de validação da internet, considerando que só é disponibilizado o conteúdo de uma lista restrita de publicações portuguesas. O processo de indexação às áreas segue os mesmos trâmites que os conteúdos de imprensa.

Para os conteúdos RTV, a CISION tem um sistema que gravação 24 horas, 7 dias por semana, em formato digital das emissões das televisões e das rádios mais relevantes para o mercado português. Posteriormente, a equipa de RTV analisa a emissão. Durante esta análise, a emissão é cortada em blocos e indexados aos clientes. Esta equipa é responsável pela produção de em média 2.500 notícias por dia.

A informação segmentada é disponibilizada no portal CISIONPOINT, como podemos visualizar na Figura 2.2. Esta plataforma foi desenvolvida internamente e foi dotada de vários módulos: Monitor, Plan, Connect e Análises.

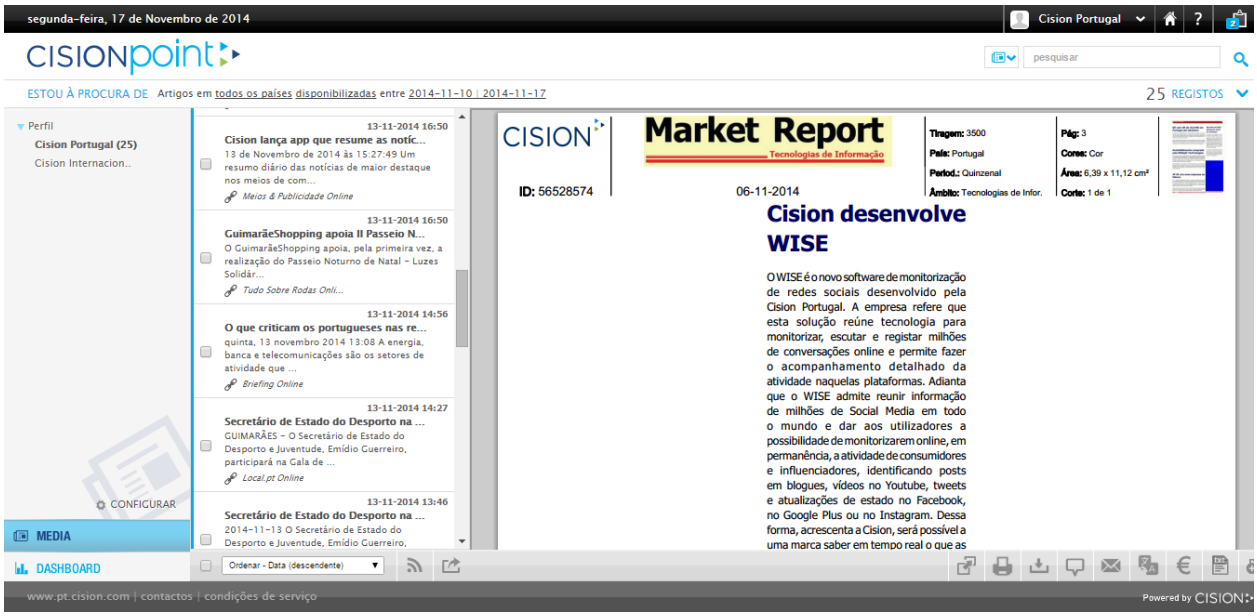


Figura 2.2 - CISIONPoint

Estas ferramentas permitem que os clientes tenham acesso à informação num curto espaço de tempo e que consigam ver e analisar a imagem que a organização/marca tem nos medias e nas redes sociais naquele momento. Atualmente, o CISIONPOINT incluiu as seguintes funcionalidades:

- Monitor:
    - Monitorização dos meios de imprensa nacional e regional, meios web, rádio e televisão;
    - Organização das notícias por áreas;
    - Pesquisa por texto, área, intervalo de datas, meios, etc;
    - Personalização do portal de acordo com a imagem do cliente;
    - "My News", onde os utilizadores organizam as notícias por pastas de acordo com os seus interesses;
    - Visualização dos artigos de imprensa e web em formato PDF ou em texto;
    - Acesso ao áudio ou ao vídeo nas notícias de rádio e televisão respetivamente;
    - Partilha da monitorização que pode ser feita através de:
        - PressBook: seleção por parte do utilizador de várias notícias dos diferentes meios e respetiva criação de um pdf;

        - Alertas: o utilizador pode definir a que horas quer receber alertas sobre uma ou mais áreas, podendo o layout ser personalizado de acordo com a imagem do cliente;
        - Email: seleção de um grupo de notícias por parte do utilizador e partilha por um conjunto destinatários via email;

- Plan:
    - Planificação de ações de comunicação;
    - Pesquisa de informação relativa a todos os meios de comunicação e autores;
    - Consulta de audiências;
    - Consulta das tabelas de publicidade;
- Connect:
    - Divulgação das ações de comunicação, planificadas no módulo "Plan";
    - Envio para diferentes contactos e plataformas;
- Análises:
    - Dashboard de artigos quantitativamente e qualitativamente.

O objeto de estudo da tese foi avaliar a possibilidade de aplicar novas técnicas de pré-indexação, com vista a melhorar as propostas enviadas para as equipas de validação de conteúdos da Internet. Atualmente, o sistema de produção está organizado por áreas e estas estão divididas por dois tipos: com ou sem validação por parte das equipas. Independentemente do seu tipo, a definição de uma área é feita através da associação de uma ou várias palavras-chave combinadas com operadores booleanos, sendo irrelevante o contexto da notícia, o que por vezes, devido à sua ambiguidade origina um grande número de propostas sem interesse. Nas áreas que incluem marcas, eventos e organizações em que o seu nome representa uma baixa ambiguidade, os artigos propostos são validados automaticamente pelo sistema.

Considerando o número de artigos validados por cada uma das equipas diariamente, verifica-se que os conteúdos da Internet é 4,2 vezes superior aos conteúdos de televisão e 2,6 vezes mais que os conteúdos de imprensa. Tendo em consideração a tendência existente por parte dos consumidores (Marktest, 2014) e por parte dos produtores de conteúdos, os artigos de internet têm tendência a aumentar, sendo preminente criar ferramentas que auxiliem e melhorem a precisão das propostas que chegam à equipa de validação da Internet. A opção de iniciar este estudo por conteúdos web, está relacionada com a exatidão com que o sistema de monitorização de conteúdos web, WISE, consegue extrair o título e o corpo da notícia, ao contrário do que acontece com o OCR na imprensa ou no sistema que faz a transcrição automática da emissão

digital dos conteúdos rádio e televisão, devido às suas especificidades. No caso concreto do OCR é o reconhecimento errado de caracteres, na televisão é a transcrição com erros.

3. Fundamentos Teóricos

O presente capítulo tem como objetivo apresentar as técnicas e metodologias utilizadas e referenciadas no desenvolvimento da solução apresentada no capítulo 4.

3.1. Contextualização

As organizações de carácter comercial quando são criadas têm como objetivo primário criar mais-valias para os seus acionistas. Estas podem faze-lo através da inovação, da eficiência, do crescimento orgânico, de fusões ou de aquisições. A gestão da informação, neste contexto, tem uma grande importância porque é a partir desta, caso a organização tenha ferramentas que lhe permita extrair e aplicar conhecimentos, que terá grande probabilidade de aumentar a sua capacidade de inovação e competitividade no mercado global. Com a evolução das tecnologias de informação (software e hardware), a quantidade de informação acumulada dentro de uma organização cresceu exponencialmente, alcançando volumes inimagináveis, não possibilitando assim a sua análise e síntese em tempo útil e não tirando por isso partido da informação recolhida sobre as atividades por si desenvolvidas. Sabe-se que no meio desta quantidade de informação, existem dados que podem ajudar no apoio à decisão sobre estratégias a implementar.

No contexto atual tudo o que acontece e existe é encontrado na Internet e caso não se encontre é porque não existe. São vários os exemplos onde uma publicação de um *post* ou de um vídeo, em poucas horas torna-se viral nas redes sociais ou em sítios de partilha de filmes como o YouTube. Um dos últimos exemplos mais mediáticos foi a *selfie* tirada por Ellen DeGeneres nos Oscars de 2014, que num espaço de doze horas foi vista por 26 milhões de vezes em toda a Web (Universe, 2014). Neste tipo de sítios os utilizadores partilham todo o género de informação, transparecendo por vezes uma certa ideia de anarquia e histeria de massas, conseguindo assim atingir uma audiência de milhões em poucas horas e influenciando, por isso, as massas.

Segundo (Universe, 2014) a produção de conteúdos digitais, ao que autor denomina de universo digital, irá apresentar um crescimento de 40% ao ano na próxima década. Para além de se incluir o aumento do número de pessoas e organizações que desenvolvem as suas atividades online, inclui-se também neste universo os dispositivos inteligentes que se ligam à Internet, desencadeando uma nova onda de oportunidades para as empresas e pessoas. O mesmo autor prevê que em 2020, o universo digital terá quase tantos bits como o número de estrelas no Universo, duplicando a cada dois anos e que chegará em 2020 aos 44 zetabytes, o que significaria ter a Lua e a Terra ligada por 6,6 pilhas de iPad Air de 128 GB.

As organizações para monitorizar todo este volume de informação necessitam de ter ferramentas quase em tempo real que lhes permita controlar/reagir a estes fenómenos num curto espaço de tempo, mas também elas próprias criarem este tipo de eventos, por forma a dar visibilidade a um produto ou a uma marca. Atualmente já existem organizações que têm departamentos especializados nesta área, de forma a melhorar a eficiência, o tempo de reação, identificar oportunidades futuras e serem especializadas em gestão de crises, quando o evento é negativo para a organização. Por estes motivos, as organizações procuram técnicas que permitam a descoberta de conhecimento em dados não estruturados, semiestruturados e estruturados, confirmando assim a ideia do autor (Berry, et al., 2010).

Para a extração de conhecimento em dados históricos estruturados, que anteriormente eram desconhecidos para a organização, são obtidos através de metodologias e técnicas de Data Mining (DM). Os resultados obtidos através destas técnicas, fornecem informações relevantes para o conhecimento do negócio, sendo assim considerado um veículo para que a organização possa criar as tais estratégias que irão garantir a sua competitividade. Segundo o autor (Thuraisingham, 1998), DM é o processo de apresentação e extração de informação relevante, padrões e de tendências, desconhecidas em grandes quantidades de dados armazenados em base de dados.

Os autores (Feldman, et al., 2007), verificam a existência de uma analogia entre o DM e o Text Mining (TM), isto porque ambos têm como objetivo a procura de informação relevante, a partir de um conjunto de dados, pela identificação e exploração de padrões. Contudo no TM, os dados são um conjunto de documentos em que a identificação dos padrões, não é feita em dados estruturados em forma de registos de base de dados, mas em dados textuais não estruturados nos documentos objeto de análise.

3.2. KDD

O DM é uma etapa de um processo maior: *Knowledge Discover in Database* (KDD). Segundo (Fayyad, et al., 1996), KDD é um processo não trivial de identificar em dados, padrões válidos, novos, potencialmente úteis e compreensíveis. Segundo o mesmo autor, o DM é uma etapa do KDD que consiste na aplicação de algoritmos específicos para a extração de padrões nos dados, enquanto o KDD refere-se a todo o processo de extração de conhecimento dos dados. Sendo KDD um processo, é necessário identificar as restantes etapas:

1. **Identificar objetivos**: identificação do domínio da aplicação e dos objetivos que o utilizador final pretende;
2. **Seleção dos dados**: identificação e seleção das fontes de dados relevantes para a extração do conhecimento;
3. **Pré-Processamento**: aplicação de operações com o objetivo de remoção dos dados inconsistentes e fora dos padrões (noise data);

4. **Transformação**: consolidação dos dados no formato apropriado para a etapa seguinte;
5. **DM**: aplicação de técnicas de análise e de extração de padrões dos dados;
6. **Interpretação\Evolução**: avaliação da informação extraída por forma a verificar se existe alguma validade para os objetivos definidos no ponto 1.

Podemos identificar as etapas referidas anteriormente, na figura seguinte.

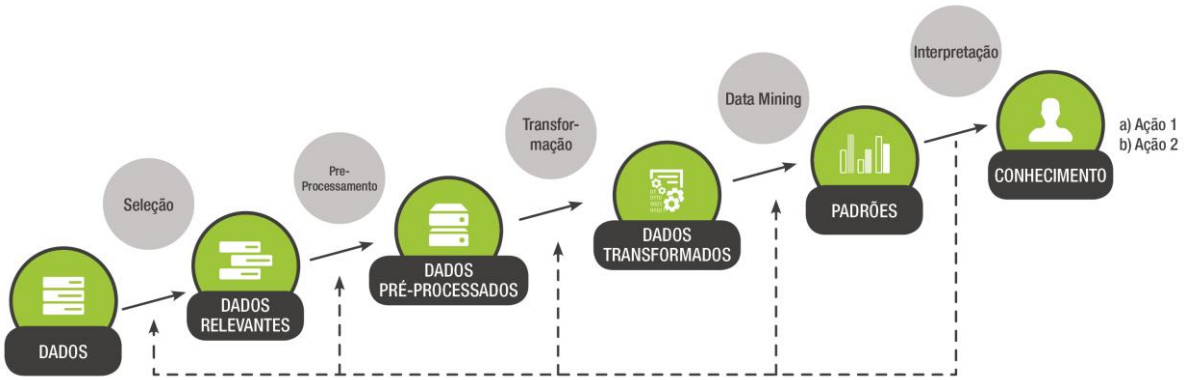


Figura 3.1 - Etapas do processo KDD (STEINER, 2006)

3.3. Text Mining

Segundo (MAHESH , et al., 2009), Text Mining (TM) é conhecido também como Text Data Mining ou a descoberta de conhecimento em base de dados textuais, referindo-se ao processo de extração de padrões interessantes e não triviais de documentos de texto não estruturados.

O TM emprega as mesmas funções analíticas que o DM, mas em dados textuais que estão organizados de uma forma não estruturada. Atualmente o grande volume de informação que é armazenado por uma organização, faz com que a sua análise seja impossível sem ajuda de sistemas ou ferramentas que façam o seu processamento automático.

Como foi referido anteriormente, os documentos de texto são dados não estruturados pelo que a extração de conhecimento torna-se difícil. O TM através de algumas técnicas tem a capacidade de transformá-los numa forma estruturada para que seja possível criar valor para as organizações através da extração do conhecimento.

Após resolução do problema da representação dos documentos numa forma estruturada, a aplicação de várias técnicas de TM permite extrair o seguinte conhecimento (Rolim, 2011):

- Sumarização: resumo de documentos;
- Categorização: agrupar documentos de acordo com o assunto;

- Clustering: agrupar documentos com caraterísticas semelhantes, independentemente da sua categorização;
- Relacionamento de Entidades: extração de entidades dos documentos e obtenção de relacionamento não óbvios entre estes;
- Resposta a questões: obtenção de respostas com base em regras de conhecimento de padrões;
- Filtragem: seleção de documentos mais relevantes para uma pesquisa ou análise;
- Routing: distribuição de documentos com base no seu conteúdo e metadados.

Para o desenvolvimento de um sistema de extração de conhecimento a partir de documentos de texto, será necessário criar um repositório que permita o armazenamento dos documentos numa forma estruturada: *document warehouse*. (Sullivan, 2001) identifica quatro atributos que caracterizam este tipo de repositório:

- Multiplicas estruturas ou tipologias de documentos;
- Múltiplas fontes;
- Extração e armazenamento automático das características e os conteúdos mais relevantes dos documentos;
- Permite a integração dos documentos através do relacionamento semântico de conteúdos.

O autor referenciado estabelece uma correspondência na caracterização dos repositórios *document warehouse*, entre as cinco principais etapas da implementação de uma data warehouse e com as deste tipo de repositório. Esta correspondência por ser verificada na Figura 3.2.

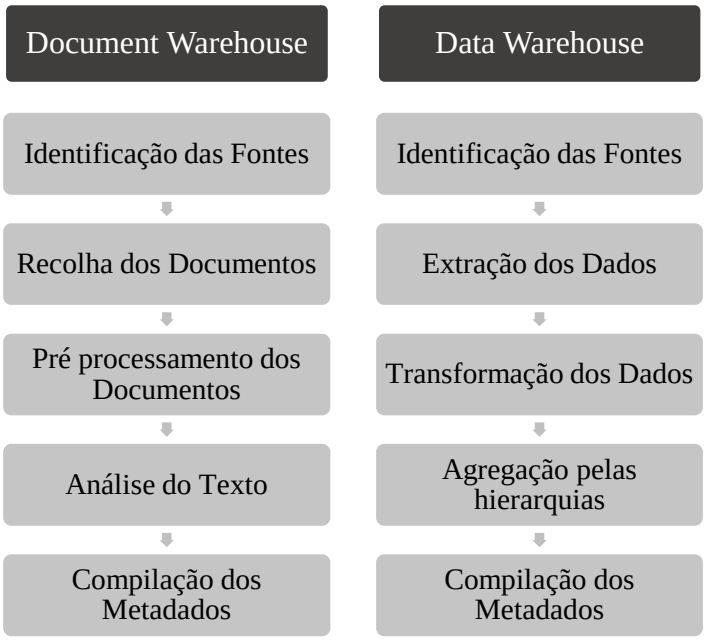


Figura 3.2 - Etapas na construção de document warehouse e data warehouse (Sullivan, 2001)

3.4. Dbpedia SpotLight

3.4.1. Reconhecimento de Entidades

Para o desenvolvimento desta tese começou-se por identificar quais as ferramentas existentes para o reconhecimento e desambiguação de entidades em documentos através de datasets públicos tais como Dbpedia e Freebase. No processo de levantamento foram consideradas as ferramentas da seguinte lista:

- OpenNLP;
- Stanford NLP;
- FreeLing;
- Apache Stanbol;
- DbPedia Spotlight.

OpenNLP

OpenNLP é uma biblioteca em Java para a realização de várias tarefas de linguagem natural de vários idiomas.

No sítio do OpenNLP já existem módulos disponíveis para *download* de modo a serem aplicados nas diversas tarefas de processamento de linguagem natural de diferentes idiomas. Por exemplo, para a língua inglesa existem módulos para as tarefas de tokenização, identificação de frases, classificador gramatical (POS Tagger) e modelos para identificação de vários tipos de entidades (nome de pessoas, organizações, localizações, etc).

Para o caso do idioma em Português, o OpenNLP apenas faculta os módulos de tokenização, detenção de frases e o classificador gramatical. Para a identificação das entidades nos textos foi necessário criar modelos de identificação com base em corpus de texto em língua portuguesa previamente classificados.

Para a língua portuguesa existe um projeto "Floresta Sintática" onde é possível encontrar vários corpus analisados morfossintaticamente e já com as entidades (pessoas, organizações, datas e etc) identificadas. Para criar os modelos, a escolha recaiu sobre a "Floresta Virgem" devido a conter textos em português europeu e brasileiro e os passos realizados foram:

1. Converter o corpus "Floresta Virgem" em formato de árvores deitadas no formato do OpenNLP:

opennlp TokenNameFinderConverter ad -encoding ISO-8859-1 -data floresatavirgem.ad -lang pt > corpus.txt

2. Criar o modelo para entidades "Person"

opennlp TokenNameFinderTrainer -lang pt -encoding UTF-8 -data corpus.txt -model pt-ner-person.bin -cutoff 20 -type person

Desde 2006 que os gestores da PT não auferiam salários e bónus como este ano, arrecadando uma quantia estimada de 12,1 milhões de euros. Contas feitas, de 2004 para cá os administradores da operadora de telecomunicações acumularam 112 milhões de euros, ao passo que a empresa desvalorizou 80%. Já os custos por colaborador sofreram uma dieta para menos 21% entre 2005 e 2013. A informação é avançada pelo jornal i.

08:24 - 03 de Setembro de 2014 | Por

Muito se tem falado da Portugal Telecom (PT), o que, nos últimos tempos, sobretudo se justifica à luz do empréstimo 'a fundo perdido' de quase 900 milhões de euros ao Grupo Espírito Santo (GES), bem como da absorção da operadora de telecomunicações portuguesa pela brasileira Oi. Esta operação saiu, aliás, prejudicada em virtude do montante cedido ao GES.

Mas vamos a contas. De acordo com a edição de hoje do jornal i, apesar deste cenário pouco auspicioso os elementos da administração da PT não viram os seus salários ou bónus minguarem. Pelo contrário. Desde 2006 que não recebiam tão avultado valor: 12,1 milhões de euros.

Se recuperarmos os montantes acumulados tomando o ano de 2004 como ponto de partida, os gestores da operadora encaixaram um total de 112,74 milhões de euros, o que equivale a mais de 10 milhões por ano.

Isto, quando no mesmo intervalo de tempo a empresa observou uma desvalorização acumulada de 78%, adianta a mesma publicação.

Mais. Entre 2005 e 2013, os custos por colaborador da PT emagreceram 21%, caindo de 38,5 mil euros para 30,4 mil euros. Ao mesmo tempo, a operadora extinguiu 1300 postos de trabalho nesse período. Estes dados, refira-se, reportam só a colaboradores da TMN e rede fixa/Meo.

Em suma, realça o i, um funcionário da PT com um vencimento médio necessitará de 30 anos para ganhar o que um administrador da empresa recebe num só ano.

Portugal subiu 15 lugares e ocupa o 36. lugar no 'ranking' mundial de competitividade de 2014-2015, divulgado hoje pelo Fórum Económico Mundial, recuperando de uma queda que se verificava desde 2005, com exceção de 2011.

O 'ranking' mundial de competitividade continua a ser liderado pela Suíça, seguida por Singapura, Estados Unidos, que subiram dois lugares, Finlândia e Alemanha, que desceram uma posição cada um, ocupam o terceiro, o quarto e o quinto lugares da tabela.

Portugal surge no 36. lugar do 'ranking', invertendo uma tendência de queda que se verificava desde 2005, quando o país alcançou o 22. lugar. O país caiu na tabela durante vários anos, à exceção de 2011, quando subiu uma posição, e no relatório divulgado no ano passado ocupou o 51. lugar.

No caso português, o Fórum destaca que "o ambicioso programa de reformas adotado pelo país parece começar a dar bons resultados", considerando, no entanto, que Portugal "não deve ser complacente e deve continuar com a implementação completa" dessas reformas, de modo a combater "as preocupações macroeconómicas persistentes".

Na verdade, o contexto macroeconómico (a dívida pública portuguesa é a 6. pior entre os 144 países analisados), o desenvolvimento do mercado financeiro e a eficiência do mercado de trabalho receberam as pontuações mais baixas no 'ranking'.

Para o Fórum, a burocracia, a carga fiscal e o acesso ao financiamento são os três fatores "mais problemáticos" para o desenvolvimento de negócios.

Entre os pontos positivos estão as infraestruturas (Portugal é o segundo país do 'ranking' com a melhor qualidade das estradas), o ensino primário e superior (as escolas de gestão ocupam o 4. lugar na tabela) e a preparação tecnológica.

A perspetiva portuguesa do relatório do Fórum Económico Mundial é apresentada esta manhã, em Lisboa pela Associação para o Desenvolvimento da Engenharia e pelo Fórum de Administradores de Empresas, num evento que conta com a intervenção do ministro da Economia, António Pires de Lima.

Figura 3.3 – Resultados da biblioteca Open NLP

Para validar os três modelos descritos anteriormente, selecionaram-se cinco textos aleatoriamente. Após a validação verificou-se que as entidades foram identificadas corretamente em geral, contudo foram identificadas falhas na caracterização do tipo de entidade. Analisando os textos da Figura 3.3 verifica-se que no texto do lado esquerdo os modelos identificaram a empresa Oi, como sendo uma pessoa.

Stanford NLP

Stanford NLP é uma biblioteca de processamento de linguagem natural em Java para o processamento de várias tarefas de vários idiomas. Para português não existia nenhum modelo treinado para identificação de entidades. Por isso foi necessário treinar o sistema utilizando o corpus "LÂMPADA 2.0" que existia para no formato aceite pela biblioteca Standford.

Freeling

A biblioteca Freeling é de alguma forma idêntica às duas anteriores, sendo que a grande vantagem é que inclui já modelos para vários idiomas: galego, asturiano, catalão, espanhol, inglês, francês, italiano e russo. Durante a fase de estudo, optou-se então por não testar esta biblioteca.

Apache Stanbol

Apache Stanbol é uma ferramenta open source que permite realizar tarefas de processamento de linguagem natural, identificação de entidades, a sua desambiguação com base em data sets públicos (dbpedia, freebse) e caso consiga realizar esta tarefa com sucesso, efetua a ligação entre a entidade no texto à sua definição, por exemplo na Wikipedia.

Por defeito, a instalação está apenas configurada para a língua inglesa, mas a arquitetura da ferramenta permite que esta seja multi-idioma. Para que tal seja possível é necessário utilizar os modelos do OPEN NLP ou Freeling. Dos testes realizados na versão online os resultados são bastante bons para a língua portuguesa.

Dbpedia Spotlight

A DbPedia Spotlight é uma ferramenta de anotação de menções de recursos da DBPedia em textos. Permite extração e desambiguação das entidades nos textos com posterior ligação à definição na DBPedia. Atualmente o sistema tem capacidade para atuar nos seguintes idiomas: inglês, alemão, holandês, francês, italiano, russo, espanhol, português, húngaro e turco.

Após a análise da documentação e realização de alguns testes na demonstração, optou-se por fazer a instalação do sistema num servidor local e consequentemente criou-se uma máquina virtual Ubuntu Server, procedendo-se da seguinte forma:

1. Fazer o download da instalação em formato debian do DBpedia Spotlight;
2. Executar o comando dpkg -i dbpedia-spotlight-0.6.deb;
3. Escolher o idioma desejado (pt);
4. Definição da quantidade de memória que deve ser alocada à DBPedia Spotlight (4G);
5. Definição do porto onde o web service da DBPedia Spotlight (8080);
6. Executar o comando /usr/bin/dbpedia-spotlight-pt.

Para a realização dos testes, utilizaram-se os cinco textos já mencionados anteriormente. Verificou-se que a identificação das entidades ocorreu dentro daquilo que era expectável.

Desta forma concluiu-se que as três primeiras bibliotecas identificavam corretamente as entidades, embora fosse necessário o desenvolvimento de um módulo que realizasse a tarefa de desambiguação, embora não tendo sido este o objeto desta tese.

Verificou-se então que os sistemas Apache Stanbol e Dbpedia realizavam as duas tarefas pretendidas: identificação e desambiguação de entidades. Comparando os dois sistemas verificou-se que o Apache Stanbol obtinha resultados com maior precisão, contudo este não

incluía os modelos de língua portuguesa para instalação. Seria então necessário utilizar uma das três bibliotecas referidas anteriormente, mas após a instalação do sistema e a tentativa de utilização das mesmas, identificou-se alguma falta de informação relativamente ao procedimento. Para a decisão da escolha da biblioteca a ser utilizada pelo Apache Stanbol seria necessário criar um corpus comum para poder optar pela melhor das três, visto que para a realização do treino para português cada uma recebe formatos diferentes. Apurados todos estes fatos, optou-se pela Dbpedia Spotlight.

3.4.2. Dbpedia Spotlight

A Dbpedia Spotlight (DS) é um sistema para anotar automaticamente documentos de texto com URIs da Dbpedia. Contém conhecimento enciclopédico da Wikipedia de cerca de 3,5 milhões de recursos, onde cerca de metade da base de conhecimento está classificado de acordo com as seguintes ontologias: pessoas, organizações ou locais. Além disso, fornece um conjunto rico em atributos e relações entre os recursos, ligando por exemplo produtos a fabricantes ou CEOs às suas empresas (Mendes, et al., 2011).

A fim de permitir a ligação de documentos Web com este hub, foi desenvolvido o Dbpedia Spotlight, um sistema que permite executar anotações em textos para os recursos da Dbpedia. Na tarefa de anotação, o utilizador fornece fragmentos de texto (documentos, parágrafos, frases) e pretende identificar URIs para os recursos mencionados. Um dos principais desafios na notação é a ambiguidade: o nome de uma entidade pode ser utilizado em diferentes contextos para referir diferentes recursos da Dbpedia. Por exemplo 'Washington' pode ser usado para referir os recursos dbpedia: George_Washington, Washington_D.C. e Washington_(US_state) entre outros. Para os leitores humanos, a desambiguação, i.e. a decisão entre candidatos para uma entidade ambígua, é normalmente realizada com base no conhecimento e no contexto de uma menção concreta (Mendes, et al., 2011).

O objectivo da Dbpedia Spotlight é proporcionar um sistema adaptável para a procura e desambiguação das menções de linguagem natural para recursos da Dbpedia. Este sistema identifica as 272 classes na ontologia da Dbpedia (Mendes, et al., 2011).

Dataset e notação

Os autores para obterem um dataset, extraem os nomes dos recursos na Dbpedia que tenham referências a outros recursos da mesma, o que os próprios denominam de *surface form*. Os recursos existentes na Dbpedia estão interligados com a Wikipedia através do nome do recurso existente nesta. Os recursos que indicam sinónimos ou outras alternativas de defenção da entidade são usados como *surface form*. De forma a completar o dataset, os autores extraem da Wikipedia todas as palavras ou expressões que contenham ligações para outras páginas da mesma, bem como os parágrafos que incluam a *surface form* para definir o contexto desta.

Os parágrafos extraídos são guardados num índice Lucene e posteriormente são segmentados, as palavras não determinantes são removidas e as palavras derivadas são reduzidas à sua forma base. No final deste processo o vetor com os termos que foi gerado é armazenado no índice Lucene para mais tarde ser usado no processo de desambiguação.

O processo inclui o cálculo da probilidade à priori para uma *surface form* corresponder a um recurso.

Para identificação das expressões candidatas a serem desimbiguadas, os autores utilizaram o LingPipe Exact Dictionary-Based Chunker que se baseia no algoritmo Aho-Corasick, o qual pesquisa palavras-chave utilizando um dicionário que contém um conjunto finito de palavras e faz diferenciação entre maíusculas e minúsculas. Os autores como dicionário utilizaram as surface forms armazenadas no dataset.

Após o processo de deteção, segue-se o processo de seleção de candidatos de forma a mapear os recursos aos candidatos e a desambiguação – usam o DBPedia Lexicalization dataset.

A próxima fase será responsável pela desambiguação da entidade, isto é, verificar qual dos candidatos é o correto. Por exemplo, se Washington se refere ao estado, cidade ou pessoa. É usado o contexto à volta dessa palavra (analisando parágrafos, links, etc.) para criar um ranking e escolher o candidato com melhor posição.

Os autores do DBpedia Spotlight usam duas métricas para obter esse ranking:

- Term Frequency (TF): este peso representa a relevância de uma palavra para um dado recurso;
- Inverse Document Frequency (IDF): este peso representa a importância de uma palavra no conjunto total de recursos disponíveis.

No entanto, o IDF não é uma boa métrica devido à natureza da associação que a palavra pode ter com o recurso que se pretende desambiguar. Por exemplo, a palavra "U.S.A" ocorre relativamente em poucos documentos, logo o seu IDF é bastante alto. Esses poucos documentos terão, à partida, a ver com o estado (que se encontra em "U.S.A"), com a cidade (capital dos "U.S.A") e com a pessoa (presidente dos "U.S.A"). Assim, o IDF deixa de conseguir dar pistas acerca da importância da palavra no contexto dos recursos a desambiguar.

Com isto em mente, os autores do DBpedia Spotlight criaram a métrica ICF (Inverse Candidate Frequency), que diz que, a relevância descritiva de uma palavra é inversamente proporcional ao número de recursos associados à mesma. Ou seja, quantos mais recursos uma palavra tem associada, menos relevância ela tem (o que se traduz num ICF mais baixo). O ICF de uma palavra w_j é calculado usando a fórmula (1).

$$ICF(w_j) = log\frac{|R_s|}{n(w_j)} log|R_s| - logn(w_j) \quad (1)$$

Onde Rs for o conjunto de recursos de uma palavra, então n(wj) é o numero de recursos em Rs associados à palavra wj (por exemplo, Washington). Usando a formula (e assumindo que |Rs| é o numero de elementos em Rs):

- se Rs = 100 e n(wj) = 2 então log(50) = 1.698 (menos relevante)
- se Rs = 100 e n(wj) = 40 então log(2,4) = 0.38 (mais relevante)

Embora não seja claro o facto de os autores do DBpedia Spotlight usarem entropia para melhorar o ranking, sabemos que o fazem porque a capacidade de descriminação de uma dada palavra é inversamente proporcional à sua entropia: uma palavra que ocorre juntamente com determinados recursos é pouco relevante para desambiguar os mesmos.

No entanto, a fórmula pode ser exemplificada usando a entropia máxima: E(w) = log n(w). Esta última é muito mais rápida e dá resultados bastante próximos aos pretendidos, pelo que o seu uso é preferível. Não sabendo como é que os autores do DBpedia Spotlight aplicam a entropia, julgo que podemos assumir que eles se limitam a usá-la para pesar o valor do ICF.

Finalmente, o ranking pode ser calculado do seguinte modo: TF×ICF. Com esta fórmula, a desambiguação para o recurso Washington é dada pela palavra que tem um valor mais alto neste ranking.

3.5. Classificação de documentos

A classificação de documentos segundo (Yang, 1999) é um problema de atribuição automática de categorias pré-definidas a documentos de texto livre.

Para além do termo de classificação de textos (*text classification*) é comum noutras publicações sobre esta problemática encontrar-se termos como categorização de textos (*text categorization*), classificação de documentos (*document classification*), categorização de documentos (*document categorization*) e descobrimento de tópicos (*topic spotting*) (Sebastiani, 2002).

Para a implementação de um sistema de classificação de documentos é necessário aplicar várias etapas até obter o classificador final. As etapas aplicadas no desenvolvimento dos classificadores de texto encontram-se representadas na Figura 3.4.

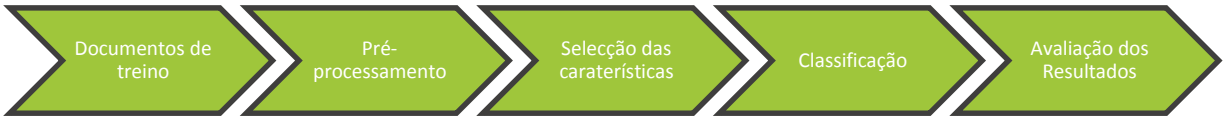


Figura 3.4 - Etapas da classificação de texto

De uma forma sucinta, podemos indicar que o processo por norma se resume às seguintes etapas (Alves, 2010):

- **Pré-Processamento**: permite tratar o documento inicial, transformando-o de forma a facilitar o processo de classificação através de métodos computacionais;
- **Classificação**: são criados classificadores que permitem atribuir categorias a documentos;
- **Avaliação dos resultados**: por fim é necessário analisar e avaliar os resultados obtidos.

3.5.1. Documentos de treino (corpus)

A primeira etapa para a implementação de um sistema de classificação de textos é a obtenção de um conjunto de treino (training corpus). Um corpus define-se como um conjunto de textos possíveis de ser interpretados pelo computador, que representam uma ou um conjunto de linguagens naturais (Gomes, 2012).

Para a obtenção do corpus empregue durante a implementação do processo explicitado no ponto 0 utilizou-se a base de dados de conteúdos da CISION Portugal referente ao segundo semestre do ano 2013. As regras para a seleção dos artigos para inclusão no corpus, encontram-se referenciados no ponto 4.1.

3.5.2. Pré-processamento

O pré-processamento pode ser definido como um processo transformativo, aplicado normalmente para reduzir o número de termos de um documento, de forma a obter uma representação mais adequada deste para as fases seguintes (Santos, 2008).

O pré-processamento tem como objetivo a transformação dos documentos, que estão numa forma não estruturada, em dados estruturados. Uma das formas mais comuns para representar um conjunto de documentos é transforma-los num vetor, onde cada elemento representa uma palavra ou termo e possui um valor numérico associado. Este valor representa a importância da palavra ou termo no conjunto dos documentos. Ao representar-se um conjunto de documentos num espaço vetorial comum, este modelo é conhecido por espaço vetorial (vetor space model). Utilizando esta forma de representação, onde a ordem das palavras ou termos não é relevante, os documentos são tratados como um saco de palavras (bag of words – BOW).

Um texto além de poder ser representado por palavras únicas (unigramas), pode ser representado por bigramas (Bekkerman, et al., 2004) ou genericamente n-grams (Caropreso, et al., 2001) podendo também representar conceitos, frases ou parágrafos. A definição de um n-gram é uma extração de n itens de uma sequência, ondes estes podem ser letras ou palavras (Santos, 2008).

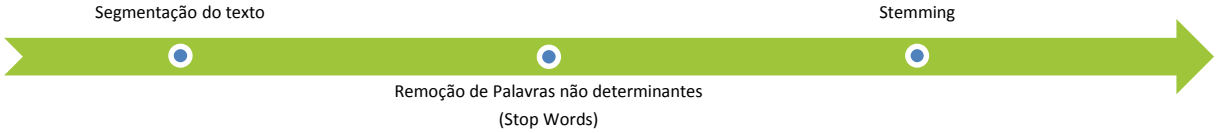


Figura 3.5 – Tarefas de Pré-Processamento

De acordo com a Figura 3.5, a primeira tarefa do pré-processamento da classificação de textos é a segmentação deste em unidades (processo conhecido como *tokenization*) obtendo assim uma estruturação dos dados (Jackson, et al., 2002). A tarefa permite ainda a remoção dos sinais de pontuação e transformação de todo o texto em minúsculas.

O passo seguinte do pré-processamento é a remoção das palavras não determinantes, ou seja, o objetivo desta fase é descartar as palavras que são muito frequentes nos documentos e que não acrescentam qualquer elemento na diferenciação entre eles. A estas palavras são chamadas de stop-words ou palavras negativas. As stop-words podem ser artigos, pronomes, preposições, advérbios entre outras palavras próprias do domínio da aplicação. As listas das stop-words variam de acordo com o idioma dos documentos.

Aplicando esta tarefa, o vetor obtido para a representação dos documentos sofre uma redução. Em vários casos o fato de não indexar stop-words não traz consequências, ou seja, palavras como: de, o, a, para, não são úteis. No entanto, se a representação dos documentos for feita através de frases, a eliminação das stop-words pode conduzir a uma perda de significado.

A última tarefa do pré-processamento tem como propósito reduzir palavras derivadas na sua forma base, ou seja, reduz-se as palavras à sua raiz semântica, removendo os afixos das palavras (sufixos e prefixos) (Spark-Jones, et al., 1997). A raiz obtida não é necessariamente igual à raiz linguística (Um estudo e apreciação sobre algoritmos de stemming para a Língua Portuguesa., 2003). O processo de stemming refere-se normalmente a um conjunto de regras heurísticas que são aplicadas para obter a raiz da palavra (Santos, 2008). Por exemplo:

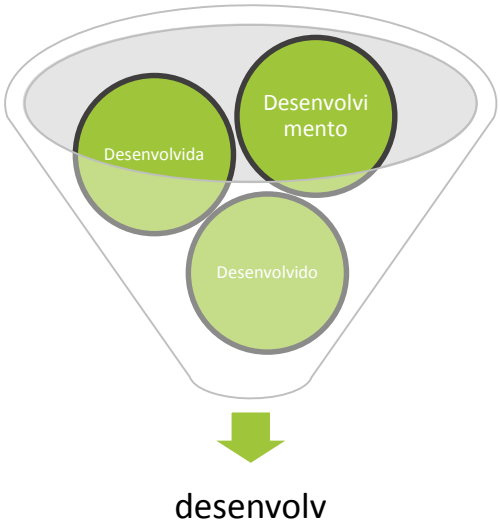


Figura 3.6 – Tarefa de Stemming (Santos, 2008)

O processo de stemming pode originar dois tipos de erro (Um estudo e apreciação sobre algoritmos de stemming para a Língua Portuguesa., 2003):

- **Overstemming**: ocorre quando é removido não apenas o sufixo, mas uma parte do *stem*. Por exemplo, a palavra gramática é transformada em grama, quando o valor correto seria gramát;
- **Understemming**: ocorre quando um sufixo não é removido completamente. Por exemplo, a palavra referência é transformada em referênc, quando o valor correto seria refer.

3.5.3. Seleção de características

A seleção de características, em inglês, feature selection, tenta identificar palavras mais relevantes dos documentos para melhorar a eficiência da categorização e reduzir a complexidade computacional (Alves, 2010) (Yang, 1999).

Durante a etapa do pré-processamento através das técnicas descritas no ponto anterior consegue-se uma redução do número de termos, contudo ainda não será suficiente para os métodos de classificação de texto. A existência de um número elevado de termos tem fortes implicações negativas na maior parte dos métodos de classificação. (Feature subset selection in text-learning, 1998) faz as seguintes observações:

- Um número elevado de termos pode diminuir significativamente o processo de aprendizagem do classificador, acabando por fornecer um resultado similar ao obtido com um conjunto de termos mais pequeno;

- Os termos usados para descrever os documentos não são necessariamente todos relevantes e benéficos para a aprendizagem do classificador, podendo mesmo reduzir a sua qualidade.

Outro meio de reduzir o conjunto de termos, que caracterizam um conjunto de documentos, é a técnica de seleção de características. Para a obtenção das características mais relevantes dos documentos procede-se da seguinte forma (Alves, 2010):

- Calcular, para cada característica, a medida que permite descriminar categorias;
- Listar as características em ordem decrescente por essa medida;
- Manter um subconjunto de características com um maior poder discriminativo possível.

Existem vários métodos possíveis para a extração das características mais importantes, mas para o desenvolvimento do BOW no ponto 4.5 foram utilizados os seguintes métodos:

- **Terms frequency (Frequência dos termos)**: calcula o número de vezes que a característica aparece num documento, só as palavras que ocorrem com mais frequência é que são guardadas. (Gonçalves, et al., 2005);
- **Chi-squared ou X^2**: utiliza a mesma tabela de contingência que a informação mútua, mas realiza um teste estatístico do X^2para inferir sobre a independência entre cada característica e categoria. A principal vantagem deste método, em comparação com a informação mútua, é que neste caso, a estatística do teste é um valor normalizado que permite comparações entre as características para a mesma categoria (Escudeiro, 2004);

3.6. Classificadores

3.6.1. Naive Bayes

Naive bayes é um classificador probabilístico baseado no Teorema de Bayes (2). O algoritmo cria um modelo que é composto por um conjunto de probabilidades.

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \quad (2)$$

Na classificação de texto, usa a probabilidade conjunta das palavras/termos e categorias, para estimar a probabilidade de uma determinada categoria de documentos. O classificador também é denominado de Naive, isto é, ingénuo, uma vez que assume que não existe relacionamento entre os atributos. Apesar disto, o classificador é simples, rápido e possui uma precisão alta e um ótimo desempenho em várias tarefas de classificação (Alves, 2010).
A probabilidade de um documento d pertencer a uma classe c é calculada através da (3).

$$P(c|d) = P(c) \prod_{1 \le k \le n_d} P(t_k|c) \quad (3)$$

Na classificação de texto, a melhor classe é obtida através do máximo à posteriori (maximum a posteriori – MAP) (4).

$$c_{MAP} = \arg max_{c \epsilon C}\ \hat{P}(c|d) = \arg max_{c \epsilon C}\ \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \qquad (4)$$

Coloca-se $\hat{P}$ para P porque desconhece-se os verdadeiros valores dos parâmetros de $P(c)$ e $P(t_k|c)$. Mas estes são estimados a partir do conjunto de treino (Manning, et al., 2009).

O classificador Naive Bayes necessita que cada probabilidade condicionada seja diferente de nula. No caso de o mesmo acontecer, utiliza-se a correção de Laplace, onde é somado 1 a todas as possibilidades.

Para demonstrar o conceito de Naive Bayes, consideremos a Figura 3.7. Os objetos podem ser classificados como verdes ou vermelhos. O objetivo é classificar um novo objeto, ou seja, se este é vermelho ou verde com base no conjunto de treino existente (Statsoft).

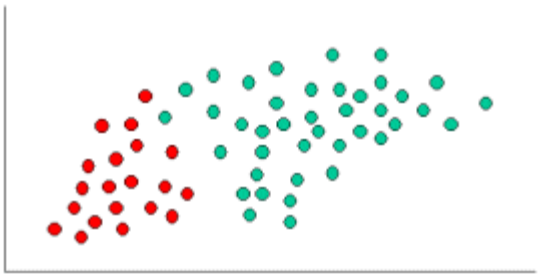


Figura 3.7 – Conjunto de treino com 2 classes (vermelho e verde)

No conjunto de treino existe duas vezes mais objetos verdes do que vermelhos, é razoável pensar que no caso de classificar um novo objeto, há duas vezes mais probabilidade de ser classificado verde do que vermelho. Neste classificador este tipo de informação é classificada com a probabilidade à priori, isto é, baseada na experiência anterior. Por isso podemos indicar que:

$$Probabilidade\ anterior_{Verde} = \frac{Número\ de\ objetos\ verdes}{Número\ total\ de\ objecto\ do\ conjunto\ de\ treino}$$

$$Probabilidade\ anterior_{Vermelho} = \frac{Número\ de\ objetos\ vermelhos}{Número\ total\ de\ objecto\ do\ conjunto\ de\ treino}$$

Sabendo que o conjunto de treino tem 60 elementos, onde 40 são verdes e 20 são vermelhos, a probalidade à priori para cada uma das classes é:

$$Probabilidade\ anterior_{Verde} = \frac{40}{60} = \frac{4}{6}$$

$$Probabilidade\ anterior_{Vermelho} = \frac{20}{60} = \frac{2}{6}$$

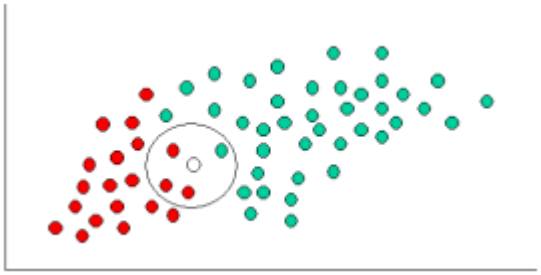


Figura 3.8 – Identificação do novo objeto a classificar

Após o cálculo da probabilidade à priori para as duas classes que compõem o conjunto de treino, pode iniciar-se a classificação de novos objetos. Considerando o objeto ($\mathcal{X}$) a branco na Figura 3.8, desenha-se um círculo em torno do objeto, de forma a englobar um número de pontos já classificados (a ser escolhido à priori) do conjunto de treino. Em seguida, conta-se o número de objetos pertencentes à classe verde e vermelho. Obtendo-se os seguintes resultados:

$$Probabilidade\ _{\mathcal{X}}\ dado\ Verde = \frac{Número\ de\ objetos\ verdes\ na\ vizinhança\ do\ \mathcal{X}}{Número\ total\ de\ objectos\ \ da\ classe\ Verde}$$

$$Probabilidade\ _{\mathcal{X}}\ dado\ Vermelho = \frac{Número\ de\ objetos\ vermelhos\ \ na\ vizinhança\ do\ \mathcal{X}}{Número\ total\ de\ objectos\ \ da\ classe\ Vermelho}$$

Considerando a Figura 3.8, verifica-se que o número de vizinhos verdes é inferior ao número de vermelhos, tendo um para três:

$$Probabilidade\ _{\mathcal{X}}\ dado\ Verde = \frac{1}{40}$$

$$Probabilidade\ _{\mathcal{X}}\ dado\ Vermelho = \frac{3}{20}$$

Embora a probabilidade à priori indique que provavelmente $\mathcal{X}$ pertence à classe Verde, apesar do número de vizinhos para este caso indicar o contrário. No classificador Naive Bayes a resposta tem em consideração estes dois valores.

$$Probabilidade\ _{\mathcal{X}}\ ser\ Verde = Probabilidade\ anterior_{Verde} \times Probabilidade\ _{\mathcal{X}}\ dado\ Verde$$
$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$Probabilidade\ _{\mathcal{X}}\ ser\ Vermelho = Probabilidade\ anterior_{Vermelho} \times Probabilidade\ _{\mathcal{X}}\ dado\ Vermelho$$
$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

O objeto $\mathcal{X}$ é classificado como Vermelho porque é a classe que tem a maior probabilidade.

Na implementação do classificador por votação no ponto 4.5.1, utilizou-se o algoritmo Naive Bayes da biblioteca NLTK em Python. A implementação do classificador por votação será abordado no ponto 4.5.1, contudo é importante referir que este algoritmo foi usado duas vezes, contudo o BWO utilizado para treinar o classificador foi criado através de metodologias diferentes. Para solucionar a possibilidade de existência de probabilidades nulas, utilizou-se em ambos a correção de Laplace.

3.6.2. Máxima Entropia

A Máxima Entropia (MaxEnt) é uma técnica para estimar a distribuição das probabilidades a partir de dados. O princípio base do algoritmo é que a distribuição deve ser tão uniforme quanto possível, onde o conjunto de treino é utilizado para obter um grupo de restrições que caracterizem uma determinada classe (Nigamy, et al., 1999).

Segundo (Nigamy, et al., 1999), este algoritmo é muito similar ao classificador Naibe Bayes. A diferença está no meio de obtenção dos parâmetros. O MaxEnt pesquisa os parâmetros que maximizem o desempenho do classificador (RATNAPARKHI, 1997), ou seja, parte do princípio que a distribuição das probabilidades deve ser o mais uniforme possível quando nada é conhecido.

Os parâmetros iniciais são aleatórios e por meio de técnicas de otimização, vão sendo refinados. Este processo garante que os parâmetros no final sejam o mais próximo possível dos valores ótimos, mas sem saber quando será conseguido. Por este motivo, o classificador poderá levar muito para aprender, no caso do conjunto de treino ter um grande número de atributos e de categorias (Gomes, 2012).

Para a implementação do classificador automático de texto por votação do ponto 4.5.1, utilizou-se a implementação do MaxEnt, com o algoritmo *GIS - Generalized Iterative Scaling* da biblioteca NLTK em Python.

3.6.3. Árvores de Decisão

As árvores de decisão (AD) são diagramas capazes de enumerar todas as probabilidades lógicas de uma sequência de decisões e ocorrências incertas. Elas mostram esquematicamente todo o conjunto de ações alternativas e acontecimentos possíveis para classificação de um novo objeto, com base num conjunto de treino (Trigueiros, 1991). As AD são compostas por (Lobo, 2010):

- Arcos: representam os resultados ou a pertença;
- Nós: representam os testes ou os conceitos;
- Folhas: são os alvos ou os conceitos terminais. Cada folha está associada a uma classe.

Como podemos verificar na Figura 3.9.

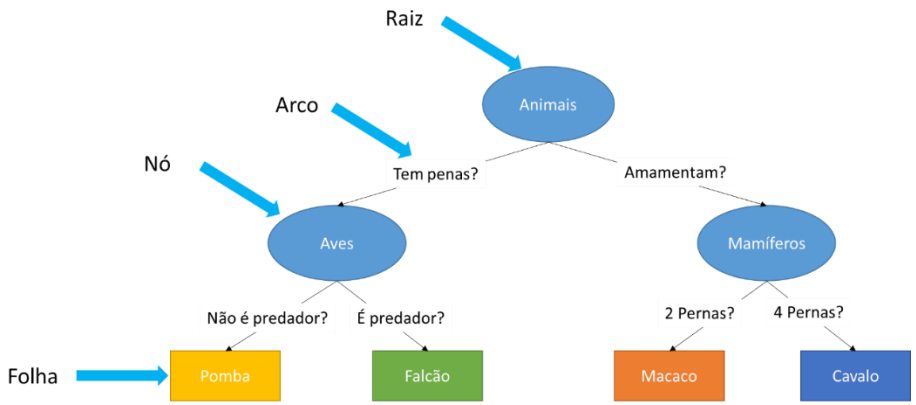


Figura 3.9 – Árvore de Decisão (Lobo, 2010)

As regras são criadas extraindo os percursos entre a raíz e as folhas.

As AD têm a vantagem de expressar o conhecimento adquirido por meio de conjuntos de regras encadeadas do tipo "SE-ENTÃO-SENÃO" ou grafo. Sendo por isso de fácil interpretação e compreensão por parte do utilizador final. Os modelos baseados em AD têm a capacidade de sumarizar grandes conjuntos de dados multivariados.

O conceito aplicado neste tipo de algoritmo é a aplicação de uma estratégia de dividir para conquistar, onde o problema complexo é decomposto em problemas mais simples, sendo que a mesma estratégia é aplicada de uma forma recursiva aos problemas já decompostos (Pereira, 2005) (Reis, et al., 2008). O processo base para a construção de uma AD é (Gama, 2002):

1. Escolher um atributo;
2. Estender a árvore adicionando um ramo para cada valor do atributo;
3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
4. Para cada folha:
    a. Se todos os exemplos são da mesma classe, associar essa classe à folha;
    b. Senão repetir os passos 1 a 4.

Existem alguns critérios que determinam o momento de paragem do processo de divisão:

- Quando todos os exemplos pertencem à mesma classe;
- Quando todos os exemplos têm os mesmos valores dos atributos de diferentes classes;
- Quando o número de exemplos é inferior a um certo limite.

Para demonstração do algoritmo, utilizaremos uma AD para decidir se um jogo de ténis se realizará ou não, com base em alguns atributos meteorológicos (MITCHELL, 1997):

- **Estado do tempo**: limpo, nublado ou chuvoso;
- **Temperatura**: alta ou baixa;
- **Humidade**: alta ou baixa;
- **Vento:** forte ou fraco.

O conjunto de treino que serve de base para criar a AD está descrito no Quadro 3.1.

Quadro 3.1 – Conjunto de treino para o jogo de ténis

Dia	Estado	Temperatura	Humidade	Vento	Jogar
D1	Limpo	Alta	Alta	Fraco	Não
D2	Limpo	Alta	Alta	Forte	Não
D3	Nublado	Alta	Alta	Fraco	Sim
D4	Chuvoso	Media	Alta	Fraco	Sim
D5	Chuvoso	Baixa	Normal	Fraco	Sim
D6	Chuvoso	Baixa	Normal	Forte	Não
D7	Nublado	Baixa	Normal	Forte	Sim
D8	Limpo	Media	Alta	Fraco	Não
D9	Limpo	Baixa	Normal	Fraco	Não
D10	Chuvoso	Media	Normal	Fraco	Sim
D11	Limpo	Media	Normal	Forte	Sim
D12	Nublado	Media	Alta	Forte	Sim
D13	Nublado	Alta	Normal	Fraco	Sim
D14	Chuvoso	Media	Alta	Forte	Não

A Figura 3.10 é a representação gráfica da árvore obtida a partir do Quadro 3.1.

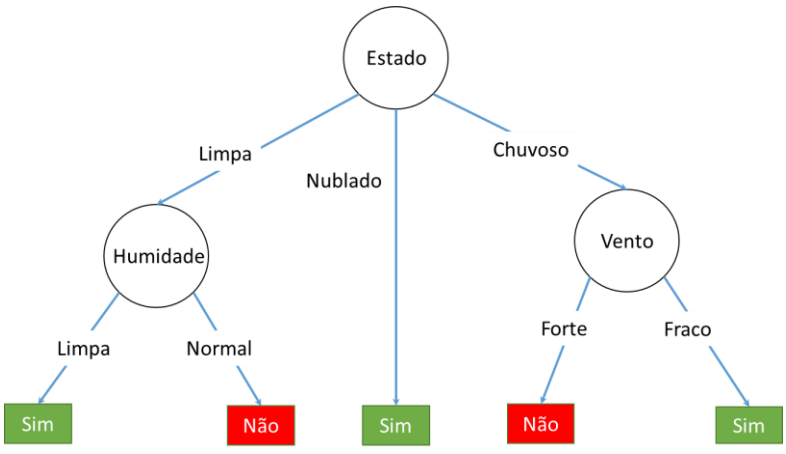


Figura 3.10 – Árvore de decisão para o exemplo do jogo ténis

Quando se decide pela implementação de uma AD, coloca-se a questão relativamente aos atributos. Como saber se um determinado atributo é elegível para descriminar as classes. Existem vários métodos, mas todos estão de acordo em relação a dois pontos (Pereira, 2005):

- Que uma divisão que mantêm as proporções iguais de classes é inútil (Figura 3.11 - a);
- Uma divisão onde em cada ramificação todos os exemplos são da mesma classe, tem utilidade máxima (Figura 3.11 - b).

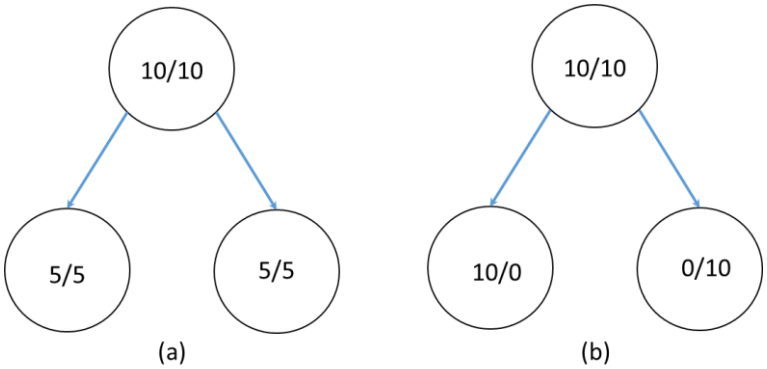


Figura 3.11 – Árvore com proporções iguais (a), Árvore com exemplos da mesma classe (b)

Entropia

A entropia de um conjunto pode ser definida como sendo o grau de pureza desse conjunto, ou seja, define a medida de falta de informação. Na Teoria da Informação este conceito existe e permite saber em média qual o número de bits necessários para representar a informação em falta, utilizando uma codificação ótima (Freitas, 2002).

Dado um conjunto S, com instâncias pertencentes à classe *i*, com probabilidade *pi*, temos:

$$Entropia(S) = \sum p_i log_2 p_i \qquad (5)$$

No exemplo do jogo de ténis apenas existem duas classes de classificação, ou seja, "Jogar Sim" (positivo, +) ou "Jogar Não" (negativo,-). Assim sendo, o valor da entropia varia de acordo com o gráfico da Figura 3.12.

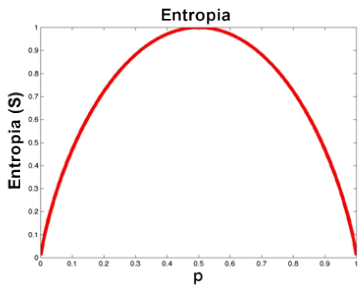


Figura 3.12 – Variação do valor da Entropia para o exemplo "Jogar Ténis" (Freitas, 2002)

Onde:

- S é o conjunto de exemplo de treino;

- p+ é a porção de exemplos positivos;
- p- é a porção de exemplos negativos;
- A entropia (6) é dada pelo desdobramento da (5).

$$Entropia(S) = -p_+ log_2 - p_- log_2 p_- \quad (6)$$

Ganho

O ganho (*gain*) define a redução esperada na entropia causada pelo particionamento do conjunto de treino por um determinado atributo, ou seja, o $Ganho(S, A)$ de um atributo A, relativamente ao conjunto de treino S, é dado pela (7) (Freitas, 2002).

$$Ganho(S, A) = Entropia(S) - \sum_{\vartheta \in Valores(A)} \frac{|S_\vartheta|}{|S|} \times Entropia(S_\vartheta) \quad (7)$$

O atributo com maior ganho de informação é obtido através da função $Máximo(Ganho(S, A))$, sabendo assim qual o atributo mais informativo (Pereira, 2005).

Exemplo "Como escolher o melhor atributo"

Usemos o exemplo do "Jogar Ténis", o primeiro passo é analisar todos os atributos, começando pela Humidade, por exemplo:

$S = [9+, 5-]$

$E = 0{,}940 = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14}$

$Ganho(S, Humidade) = 0{,}151$

$S_{humidade,elevada} = [3+, 4-]$ $S_{humidade,normal} = [6+, 1-]$

$E_{humidade,elevada} = 0,985$ $E_{humidade,normal} = 0,592$

$Ganho(S, Humidade) = 0{,}940 - \frac{7}{14} \times 0{,}985 - \frac{7}{14} \times 0{,}592 = 0{,}151$

Calculo do ganho para o atributo **Estado**:

$Ganho(S, Estado) = 0{,}247$

$S_{estado,limpo} = [2+, 3-]$ $S_{estado,nublado} = [6+, 1-]$ $S_{aspeto,chuvoso} = [3+, 2-]$

$E_{estado,limpo} = 0,971$ $E_{estado,nublado} = 0$ $E_{estado,chuvoso} = 0{,}971$

$Ganho(S, Estado) = 0{,}940 - \frac{5}{14} \times 0{,}971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0{,}0971 = 0{,}247$

Calculo do ganho para o atributo **Vento**:

$Ganho(S, Vento) = 0{,}048$
$S_{vento,fraco} = [6+, 2-] \qquad S_{vento,forte} = [3+, 3-]$
$E_{vento,fraco} = 0,811 \qquad E_{vento,forte} = 1$
$$Ganho(S, Vento) = 0{,}940 - \frac{8}{14} \times 0{,}811 - \frac{6}{14} \times 1 = 0{,}048$$

Calculo do ganho para o atributo **Temperatura**:

$Ganho(S, Temperatura) = 0{,}029$
$S_{temperatura,alta} = [2+, 2-]\ S_{temperatura,media} = [4+, 2-]\ S_{temperatura,baixa} = [3+, 1-]$
$E_{temperatura,alta} = 1 \qquad E_{temperatura,media} = 0{,}918 \qquad E_{temperatura,baixa} = 0{,}811$
$$Ganho(S, Temperatura) = 0{,}940 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0{,}918 - \frac{4}{14} \times 0{,}811 = 0{,}048$$

Aplicando a fórmula do $Máximo(Ganho(S, A))$, obtemos:

$$Máximo \begin{pmatrix} Ganho(S, Humidade) \\ Ganho(S, Vento) \\ Ganho(S, Estado) \\ Ganho(S, Temperatura) \end{pmatrix} = Ganho(S, Estado)$$

Como se pode verificar, o atributo que fornece maior ganho é o atributo Estado, sendo o escolhido para a raiz da árvore, confirmando assim a Figura 3.10.

O próximo passo será calcular os ganhos obtidos para cada uma das três situações possíveis:

$Chuvoso = [2+, 3-] \quad E > 0 \quad Classe???$
$Nublado = [4+, 3-] \quad E = 0 \quad JogarSim$
$Limpo = [3+, 2-] \quad E > 0 \quad Classe???$

Na situação do atributo **Estado** ser igual a *Nublado* fica identificado a classe, uma vez que a totalidade dos registos tem a mesma classe, $JogarSim$.

Para obter a estrutura da AD esquematizada na Figura 3.10, os restantes cálculos são efetuados de forma análoga aos demonstrados.

Na implementação do classificador automático de texto por votação do ponto 4.5.1, utilizou-se a biblioteca NLTK em Python. Um dos classificadores utilizados para a votação foi uma AD.

A biblioteca NLTK utiliza o algoritmo Decision Stump no classificador *nltk.classify.decisiontree*. O algoritmo consiste em criar uma AD com apenas um nível e duas

ramificações. Normalmente este tipo de algoritmo não é aplicado sozinho, mas sim em sistemas de classificação por votação. Esta situação verifica-se pelo fato de se usar apenas um atributo, tornando por isso o seu desempenho baixo. O algoritmo encontra a melhor divisão dos atributos no conjunto de treino através da maximização da entropia (Duarte, 2009).

Um sistema de classificação por votação tem como objetivo melhorar o desempenho do mesmo, ou seja, diminuir o número de falsos positivos, aumentando assim a precisão. Estes sistemas combinam a saída de vários classificadores, a classe vencedora será aquela que obtiver a maioria dos votos. O número de classificadores deverá ser impar para evitar empates.

3.6.4. KNN

O conceito do algoritmo *k-nearest neighbors* (KNN) é um algoritmo de fácil implementação e compreensão. O método não necessita de uma fase de aprendizagem, apenas faz o cálculo de uma métrica e a estes métodos denominam-se de *lazy learning* (Teixeira, 2014). O algoritmo procura um K de vizinhos mais similares ao objeto em análise no conjunto de treino e a classe atribuída é aquela que predomina nos vizinhos destes (X. Wu, 2007).

Dado um conjunto de treino $\mathcal{D}$ descrito por $x = (\mathbf{x}', y')$ formado por padrões de entrada (x') e pela respetiva etiqueta (y'), calcula-se a distância entre um determinado padrão z e todos os padrões existentes em $\mathcal{D}$, para depois se obter os K vizinhos mais próximos. Desta forma é possível definir um subconjunto de $\mathcal{D}$ chamado que engloba os K padrões mais próximos de z (Teixeira, 2014) (X. Wu, 2007). Após a lista de vizinhos mais próximos ter sido obtida, o z é classificado com base na classe em maioria na lista obtida (X. Wu, 2007).

$$y' = argmax_v \sum_{(\mathbf{x_i}, \mathbf{y_i}) \epsilon\, D_z} I(v = y_i) \tag{8}$$

A (8) é a representação matemática para a votação por maioria (David Reby, 1997), onde v é uma etiqueta da classe, y_i é a etiqueta das classes para os vizinhos mais próximos e a função $I(v = y_i)$ devolve 1, no caso de $v = y_i$ ou devolve 0, no caso contrário (Teixeira, 2014) (X. Wu, 2007).

O algoritmo pode ser utilizado para classificações e em análises de regressões. O resultado obtido depende do tipo de análise (algorithm, 2014):

- Na classificação, o resultado é a classe em maioria nos vizinhos do objeto a ser classificado;
- Na regressão o resultado é obtido através da média dos valores dos seus vizinhos mais próximos.

Consideremos o seguinte exemplo (Yhat, 2013), temos um conjunto de valores numéricos que estão classificados de acordo com as classes "Azul" ou "Vermelho". Cada objeto é caracterizado por 2 valores: x e y (Gráfico 3.1).

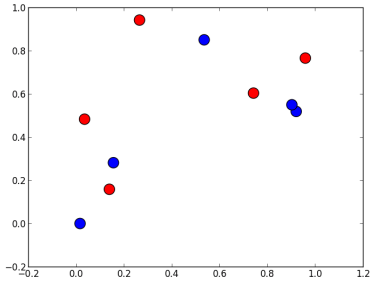


Gráfico 3.1 – Disposição gráfica dos objetos

Entretanto surge um novo objeto com as caraterísticas x_1 e y_1 para ser classificado como "Azul" ou "Vermelho", sendo representado no Gráfico 3.2 através de um ponto preto.

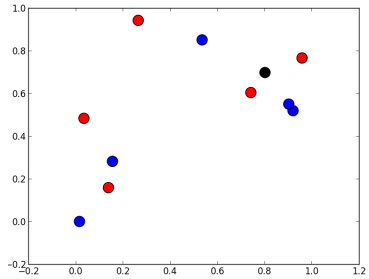


Gráfico 3.2 – Representação gráfica do novo objeto a preto

O algoritmo KNN tem como abordagem obter os objetos mais próximos do objeto por classificar. Para este caso, consideremos um $K = 3$, ou seja, selecionar os 3 vizinhos mais próximos do novo objeto, como sugere o Gráfico 3.3. Os vizinhos selecionados votam e a cor que for mais predominante é atribuída ao objeto a preto, sendo a cor predominante o vermelho, como poderemos verificar no Gráfico 3.4.

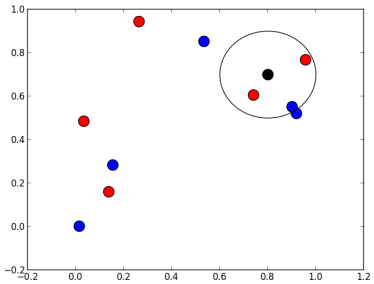


Gráfico 3.3 - Seleção dos 3 vizinhos mais próximos do ponto a preto

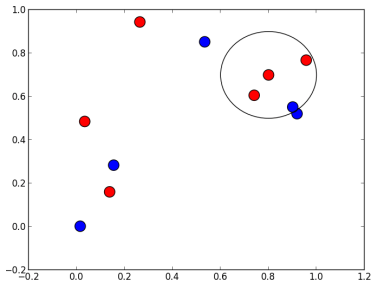


Gráfico 3.4 – Resultado da votação por parte dos vizinhos

A dificuldade que este algoritmo apresenta é a definição do número de vizinhos a selecionar. Segundo (Hsinchun Chen, 2006), escolher o melhor valor para K não é fácil, mas trabalhoso. Uma premissa que deverá ser respeitada na implementação deste algoritmo é que o K deverá ser impar para evitar que existam empates no resultado da votação.

3.6.5. Métricas

Na implementação de um modelo de classificação há necessidade de saber qual a sua a capacidade de generalização. Para validação de um modelo existem várias métricas nos permitem perceber se o modelo criado é bom ou não. Durante a fase de implementação das duas abordagens propostas para a classificação de documentos, nos pontos 4.4 e 4.5.1, foram utilizadas as seguintes métricas: *precision*, *recall* e *accurracy*.

As métricas para serem calculadas necessitam da matriz de confusão (Quadro 3.2), onde os diferentes tipos de erro e acertos são sintetizados (Pereira, 2005).

Quadro 3.2 – Matriz Confusão Genérica (Pereira, 2005)

	Predição Positiva	**Predição Negativa**
Classe Positiva	Verdadeiro Positivo (TP)	False Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Onde:

TP – True Positive (Positivos Verdadeiros) FN – False Negative (Negativos Falsos)
FP – False Positive (Positivos Falsos) TN – True Negative (Negativos Verdadeiros)

O quadro seguinte resume as métricas mencionadas anteriormente.

Quadro 3.3 – Resumo das Métricas

Métrica	Descrição	Equação
Accuracy	Qual a percentagem de previsões que está correta	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Qual a percentagem de documentos relevantes	$\frac{TP}{TP + FP}$
Recall	Qual a percentagem de documentos relevantes que foram recuperados	$\frac{TN}{TP + FP}$

3.7. Similaridade de documentos

No ponto anterior abordou-se a classificação de documentos através da utilização de classificadores probabilísticos. Outro método também utilizado foi a classificação através do grau de similaridade entre documentos. Para a resolução do problema identificado no ponto 4.3.3, optou-se pelo método referido anteriormente.

Na implementação criou-se um índice do tipo *Latent Semantic Indexing* (LSI) para indexação dos documentos com o objetivo de medir o grau de similaridade entre estes e os novos documentos. Como unidade de medida utilizou-se o cosseno do ângulo entre dois vetores (9).

$$\cos(x, y) = \frac{x \times y}{\|x\| \times \|y\|} \tag{9}$$

Esta medida é invariante com a escala, isto é, não depende do tamanho dos vetores, mas apenas da sua direção. Assim permite tratar de modo idêntico documentos com a mesma distribuição relativa de termos (Trigo, 2010).

A indexação semântica latente, do inglês *Latent Semantic Indexing* (LSI), é um método de indexação e recuperação de informação que usa a técnica matemática Singular Value Decomposition (SVD) para encontrar padrões nas relações entre os termos e os conceitos incluídos num conjunto de textos não estruturados. O LSI é baseado no princípio de que as palavras que são usadas no mesmo contexto tendem a ter significados semelhantes. A característica essencial do LSI é a sua capacidade de extrair o conteúdo conceptual do corpo do texto através da criação de associações entre os termos que ocorrem em contextos similares (Wikipedia).

O LSI transforma um conjunto de documento numa matriz, denominada de termo-documento. Onde nas colunas estão documentos e nas linhas estão os termos indexados de cada documento. Por exemplo, seja t_i a linha e d_j a coluna, o A_{ij} é elemento da matriz que representa o número de vezes que o termo i aparece no documento j. À matriz anteriormente referida, é aplicada a operação Simple Value Decomposition (SVD), para efetuar a divisão desta em três matrizes:

- Matriz U que contém os termos;
- Matriz S que contém os valores mais representativos da matriz termo-documento (os valores singulares);
- Matriz V que contém os documentos.

Após a criação destas três matrizes, é escolhido um tamanho, K, para criar três novas matrizes (que serão chamadas U', S' e V'). A estas três novas matrizes é multiplicado o vetor Q, que representa uma pesquisa. O resultado desta multiplicação será um vetor cujo conteúdo é uma lista dos documentos mais relevantes para a pesquisa requisitada.

Para a implementação do processo descrito anteriormente utilizou-se a biblioteca Gensim em Python.

4. Apresentação de resultados

No presente capítulo descreve-se todo o desenvolvimento do projeto, a forma como decorreu a investigação e os respetivos resultados finais.

4.1. Dados

No decorrer da investigação, apenas foram considerados artigos da internet, tendo sido selecionados aleatoriamente seis intervalos de datas do 2º semestre do ano de 2013, obtendo-se em média 55 mil artigos por cada intervalo. Deste conjunto foram excluídos os sítios classificados como blogs na base de medias da CISION Portugal. Esta exclusão deve-se ao facto de este tipo de informação não ter relevância para algumas áreas. Os intervalos de datas selecionados foram:

- 05-08 A 12-08;
- 13-08 A 20-08;
- 23-09 A 30-09;
- 07-10 A 14-10;
- 18-11 A 25-11;
- 09-12 A 16-12.

Para que o conjunto de dados ficasse completo, definiu-se dois critérios para a seleção das áreas:

1. **Validação automática**: áreas que não necessitam de validação por parte das equipas de produção;
2. **Baixa ambiguidade**: as palavras-chave que definem a área não deixam dúvidas quanto à interpretação num determinado contexto.

Aplicando os critérios definidos foram selecionadas as seguintes áreas:

- Transportes Aéreos de Portugueses (TAP);
- Portugal Telecom (PT);
- Pirelli;
- Banco de Portugal (BdP);
- Air France;
- Cristiano Ronaldo (CR7).

4.2. Configurações

No sistema que se encontra em funcionamento na CISION Portugal, as áreas existentes são definidas pela combinação booleana de uma ou várias palavras-chave, estando estas definições armazenadas numa base de dados relacional de um servidor SQL Server. Para o armazenamento das configurações e das indexações dos artigos às áreas desta investigação, seguiu a mesma abordagem. Procedeu-se ao levantamento dos requisitos e identificou-se as seguintes tabelas:

- Area (Quadro 4.1);
- AreaSearchProfile (Quadro 4.2);
- Article (Quadro 4.3)
- ArticleArea (Quadro 4.4).

Quadro 4.1 – Area: Informação relativa à área

Campo	Tipo	Descrição
AreaId	numeric(18)	PK da tabela
Name	nvarchar(150)	Nome da área
Metodo	Int	Método de desambiguação a ser aplicado

Quadro 4.2 – AreaSearchProfile: filtros associados à área

Campo	Tipo	Descrição
AreaSearchProfileId	numeric(18)	PK da tabela
Filter	nvarchar(150)	Filtro a ser aplicado na pesquisa
AreaId	numeric(18)	FK da tabela Area

Quadro 4.3 - Article: Artigos captados pelo sistema WISE

Campo	Tipo	Descrição
ArticleId	numeric(18)	PK da tabela
Headline	nvarchar(300)	Título do artigo
BodyText	nvarchar(max)	Texto do artigo
MediaName	nvarchar(150)	Nome da publicação
PublicationDate	smalldatetime	Data de Publicação
Link	nvarchar(300)	Url do artigo
Status	Int	Estado do artigo 0 – Nova indexação por parte do WISE 1 – Fase 1 do processo finalizada 2 – Fase 2 do processo finalizada

Quadro 4.4 – ArticleArea: Artigos associados a uma área

Campo	Tipo	Descrição
ArticleId	numeric(18)	PK da tabela (chave composta) FK da tabela Article
AreaId	numeric(18)	PK da tabela (chave composta) FK da tabela Area
AreaSearchProfileId	numeric(18)	FK da tabela AreaSearchProfile

Após a especificação da estrutura de base de dados, implementou-se num servidor SQL Server como se pode verificar na Figura 4.1:

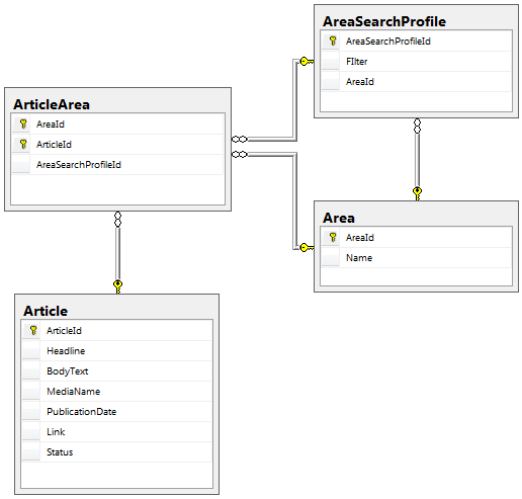


Figura 4.1 – Diagrama de base de dados de apoio às áreas

As áreas selecionadas no ponto 4.1 foram configuradas de acordo com as seguintes regras:

1. Cada área pode ser definida por mais do que uma entidade, sendo que estas podem ser combinadas com os operadores booleanos;
2. Não poderá existir pesquisas através de palavras-chave.

No Quadro 4.5 é possível comparar o método atual e o proposto nesta investigação, para a definição de uma área. O exemplo utilizado é referente à área TAP.

Quadro 4.5 – Palavras-chave vs Entidade da área TAP

Produção	Investigação
TAP OR Transportes Aéreos Portugueses	TAP_Portugal

Procedeu-se à definição das áreas selecionadas no ponto 4.1 através de entidades da DbPedia de acordo com o Quadro 4.6.

Quadro 4.6 – Definição das áreas através de entidades da DbPedia

Área	Definição
Transportes Aéreos de Portugueses (TAP)	TAP_Portugal
Portugal Telecom (PT)	Portugal_Telecom
Pirelli	Pirelli
Banco de Portugal (BdP)	Banco_de_Portugal
Air France	Air_France
Cristiano Ronaldo	Cristiano_Ronaldo

4.3. Sistema – Fase 1

O processo de indexação de novos artigos a áreas, com base em reconhecimento de entidades, foi dividido em duas fases:

- Anotação e desambiguação de entidades através da utilização da Dbpedia Spotlight (DS);
- Indexação dos conteúdos a áreas através da combinação booleana de uma ou mais entidades.

4.3.1. Anotação e Desambiguação através da Dbpedia Spotlight

Nos últimos anos a CISION Portugal armazena diariamente dois milhões de artigos do tipo internet e sobre os quais aplica milhares de pesquisas, através da utilização de um índice num servidor SOLR. Face a esta experiência adquirida, os sistemas similares devem incluir um índice deste género.

Os artigos extraídos pelo sistema WISE são numa primeira fase armazenados numa base de dados relacional, sendo posteriormente indexados a um índice SOLR com a estrutura definida no Quadro 4.7.

Quadro 4.7 – Campos do Índice SOLR

Campo	Descrição
Id	Chave primária de cada registo. Igual ao ID do artigo na base de dados de produção
Headline	Título da notícia
BodyText	Corpo da notícia
PublicationDate	Data de publicação
SiteName	Nome do mídia
EntitiesSpotlight	Lista de entidades reconhecidas pelo DbPediaSpotlight. Este campo é multivalor, funcionando como array.
EntitySpotlight_*	Informação relativa a cada entidade extraída pela DbPedia Spotlight. Este sistema devolve a seguinte informação: • **Name**: nome da entidade existente na Wikipedia; • **Support**: número de ligações que a definição tem na dbpedia; • **URL**: link para a Dbpedia; • **SurfaceForm**: palavra ou expressão que foi relacionada com a entidade; • **SimilirityScore**: qual o grau de certeza da proposta

O segundo passo foi criar um serviço Windows que fosse responsável pela obtenção de novos artigos na base de dados (indexados previamente pelo WISE) e enviar cada um deles para o sistema DS. Este serviço faz a anotação e a desambiguação das entidades reconhecidas no texto, para a sua posterior indexação no índice SOLR, conforme a estrutura descrita no Quadro 4.7 apresentado na página anterior.

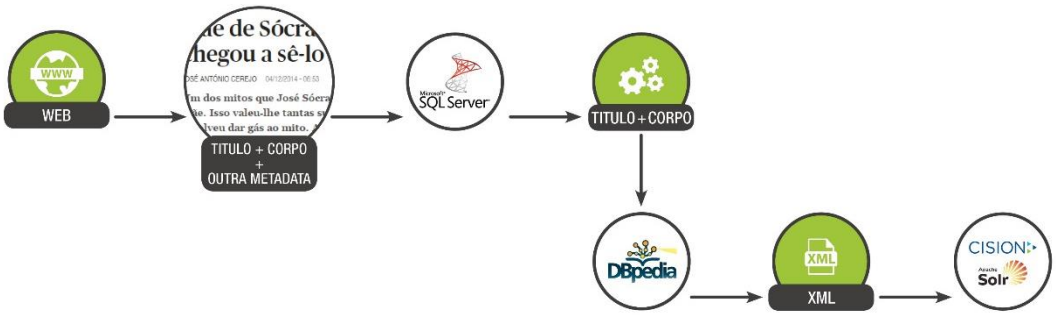
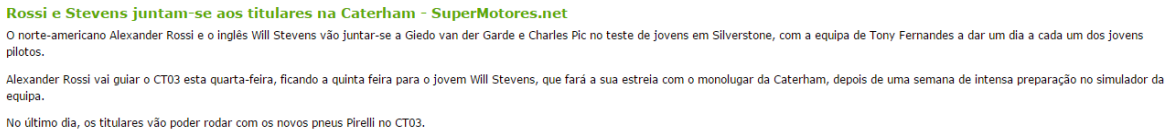


Figura 4.2 – Fase 1 do processo de indexação com base no reconhecimento de entidades

Consideremos o seguinte exemplo para demonstração da fase 1 do processo que se encontra esquematizado na Figura 4.2.

Suponhamos que o próximo sítio a ser verificado pelo WISE para identificação de novos artigos é o SuperMotores.net. O WISE extrai todos os links que correspondem a notícias e destes os que são novos desde a última verificação. Consideremos que durante o processo de análise, se identificou apenas um novo link, procedeu-se ao download da página para a extração do título, do corpo da noticia (ver Figura 4.3), do link, da data de publicação para o seu armazenamento na tabela Article com o campo Status igual a um. O artigo após este processo fica então disponível para anotação e desambiguação de entidades.



Rossi e Stevens juntam-se aos titulares na Caterham - SuperMotores.net

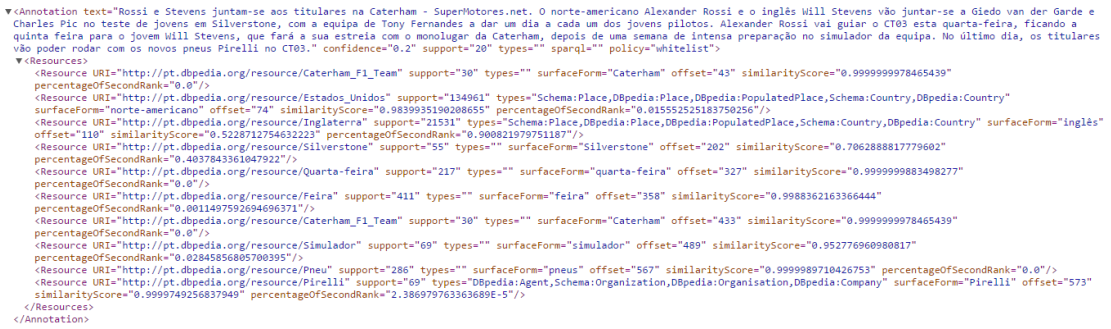
O norte-americano Alexander Rossi e o inglês Will Stevens vão juntar-se a Giedo van der Garde e Charles Pic no teste de jovens em Silverstone, com a equipa de Tony Fernandes a dar um dia a cada um dos jovens pilotos.

Alexander Rossi vai guiar o CT03 esta quarta-feira, ficando a quinta feira para o jovem Will Stevens, que fará a sua estreia com o monolugar da Caterham, depois de uma semana de intensa preparação no simulador da equipa.

No último dia, os titulares vão poder rodar com os novos pneus Pirelli no CT03.

Figura 4.3 – Título e corpo da notícia de um artigo do sítio SuperMotores.net

Na fase 2, o serviço Windows após a última verificação, identifica o artigo publicado na SuperMotores.net e coloca-o em fila de espera para o processo de anotação e desambiguação de entidades. O serviço obtém o título e o corpo do artigo (ver Figura 4.3), faz a concatenação dos dois textos num só, envia para a DS e recebe de volta o texto anotado. A resposta enviada pela DS é recebida em formato de XML (ver Figura 4.4).



```
▼<Annotation text="Rossi e Stevens juntam-se aos titulares na Caterham - SuperMotores.net. O norte-americano Alexander Rossi e o inglês Will Stevens vão juntar-se a Giedo van der Garde e
 Charles Pic no teste de jovens em Silverstone, com a equipa de Tony Fernandes a dar um dia a cada um dos jovens pilotos. Alexander Rossi vai guiar o CT03 esta quarta-feira, ficando a
 quinta feira para o jovem Will Stevens, que fará a sua estreia com o monolugar da Caterham, depois de uma semana de intensa preparação no simulador da equipa. No último dia, os titulares
 vão poder rodar com os novos pneus Pirelli no CT03." confidence="0.2" support="20" types="" sparql="" policy="whitelist">
 ▼<Resources>
    <Resource URI="http://pt.dbpedia.org/resource/Caterham_F1_Team" support="30" types="" surfaceForm="Caterham" offset="43" similarityScore="0.9999999978465439"
    percentageOfSecondRank="0.0"/>
    <Resource URI="http://pt.dbpedia.org/resource/Estados_Unidos" support="134961" types="Schema:Place,DBpedia:Place,DBpedia:PopulatedPlace,Schema:Country,DBpedia:Country"
    surfaceForm="norte-americano" offset="74" similarityScore="0.9839935190208655" percentageOfSecondRank="0.015552525183750256"/>
    <Resource URI="http://pt.dbpedia.org/resource/Inglaterra" support="21531" types="Schema:Place,DBpedia:Place,DBpedia:PopulatedPlace,Schema:Country,DBpedia:Country" surfaceForm="inglês"
    offset="110" similarityScore="0.5228712754632223" percentageOfSecondRank="0.900821979751187"/>
    <Resource URI="http://pt.dbpedia.org/resource/Silverstone" support="55" types="" surfaceForm="Silverstone" offset="202" similarityScore="0.7062888817779602"
    percentageOfSecondRank="0.4037843361047922"/>
    <Resource URI="http://pt.dbpedia.org/resource/Quarta-feira" support="217" types="" surfaceForm="quarta-feira" offset="327" similarityScore="0.9999999883498277"
    percentageOfSecondRank="0.0"/>
    <Resource URI="http://pt.dbpedia.org/resource/Feira" support="411" types="" surfaceForm="feira" offset="358" similarityScore="0.998836216336444"
    percentageOfSecondRank="0.0011497592694696371"/>
    <Resource URI="http://pt.dbpedia.org/resource/Caterham_F1_Team" support="30" types="" surfaceForm="Caterham" offset="433" similarityScore="0.9999999978465439"
    percentageOfSecondRank="0.0"/>
    <Resource URI="http://pt.dbpedia.org/resource/Simulador" support="69" types="" surfaceForm="simulador" offset="489" similarityScore="0.952776960980817"
    percentageOfSecondRank="0.02845856805700395"/>
    <Resource URI="http://pt.dbpedia.org/resource/Pneu" support="286" types="" surfaceForm="pneus" offset="567" similarityScore="0.9999989710426753" percentageOfSecondRank="0.0"/>
    <Resource URI="http://pt.dbpedia.org/resource/Pirelli" support="69" types="DBpedia:Agent,Schema:Organization,DBpedia:Organisation,DBpedia:Company" surfaceForm="Pirelli" offset="573"
    similarityScore="0.9999749256837949" percentageOfSecondRank="2.386979763363689E-5"/>
  </Resources>
</Annotation>
```

Figura 4.4 – Texto anotado pela DS em formato XML

Figura 4.5 – Texto anotado pela DBSL via interface web

Para a visualização gráfica do resultado em formato XML, em situações pontuais, utilizou-se o interface web disponibilizado pelo DS (ver Figura 4.5). Utiliza-se este interface neste exemplo, para simplificar a análise, para que seja possível visualizar as expressões do texto anotadas e desambiguadas para entidades na DBpedia. As expressões sublinhadas no texto, os autores da DS classificam-nas como surfaceform. Por exemplo:

- *surfaceform* **Caterham** é a entidade na DbPedia:Caterham_F1_Team, com um grau de similaridade de 99%;
- *surfaceform* **norte-american**o é a entidade na DbPedia:Estados_Unidos, com um grau de similaridade de 98%.

Após a obtenção do resultado, o serviço Windows procede ao parse do XML e ao respetivo armazenamento no índice SOLR. O campo Status, do registo do artigo na tabela *Article*, é modificado para dois. Este estado indica que o artigo encontra-se disponível para ser indexado a áreas. A estrutura do artigo no índice SOLR pode ser visualizada na figura seguinte.


```
Id: "48778768",
Headline: "Rossi e Stevens juntam-se aos titulares na Caterham » SuperMotores.net",
BodyText: " em / por / em 16 de Julho de 2013 às 20:30 / O norte-americano Alexander Rossi e o inglês Will Stevens vão juntar-se a Giedo van der Garde e Charles Pic no teste de jovens em
Silverstone, com a equipa de Tony Fernandes a dar um dia a cada um dos jovens pilotos. Alexander Rossi vai guiar o CT03 esta quarta-feira, ficando a quinta feira para o jovem Will Stevens,
que fará a sua estreia com o monolugar da Caterham, depois de uma semana de intensa preparação no simulador da equipa. No último dia, os titulares vão poder rodar com os novos pneus Pirelli
no CT03. Fórmula 1 Ricardo Batista ",
PublicationDate: "2013-07-17T00:00:00Z",
SiteName: "Super Motores.net",
- EntitiesSpotlight: [
    "Caterham_F1_Team",
    "16_de_julho",
    "Estados_Unidos",
    "Inglaterra",
    "Circuito_de_Silverstone",
    "Quarta-feira",
    "Feira",
    "Simulador",
    "Pneu",
    "Pirelli",
    "Fórmula_1"
],
EntitySpotlight_1094132931-Name: "Caterham_F1_Team",
EntitySpotlight_1094132931-Support: "30",
EntitySpotlight_1094132931-Url: "http://pt.dbpedia.org/resource/Caterham_F1_Team",
EntitySpotlight_1094132931-SurfaceForm: "Caterham",
EntitySpotlight_1094132931-SimilarityScore: "0,9999999999373586",
EntitySpotlight_-1118454480-Name: "16_de_julho",
EntitySpotlight_-1118454480-Support: "1017",
EntitySpotlight_-1118454480-Url: "http://pt.dbpedia.org/resource/16_de_julho",
EntitySpotlight_-1118454480-SurfaceForm: "16 de Julho",
EntitySpotlight_-1118454480-SimilarityScore: "1,0",
EntitySpotlight_-1947831162-Name: "Estados_Unidos",
EntitySpotlight_-1947831162-Support: "134961",
EntitySpotlight_-1947831162-Url: "http://pt.dbpedia.org/resource/Estados_Unidos",
EntitySpotlight_-1947831162-Type: "Schema:Place,DBpedia:Place,DBpedia:PopulatedPlace,Schema:Country,DBpedia:Country",
EntitySpotlight_-1947831162-SurfaceForm: "norte-americano",
EntitySpotlight_-1947831162-SimilarityScore: "0,9754252109486307",
```


Figura 4.6 – Artigo com as entidades reconhecidas armazenado no índice SOLR

Podemos visualizar na Figura 4.6, o identificador único do artigo, o título, o corpo da notícia, a data publicação, o sítio, a lista de entidades e a resposta devolvida pelo DS.

Com o objetivo de facilitar a integração do DS em processos externos, os autores do sistema disponibilizaram um web service com métodos que permitem fazer o processo de anotação e desambiguação de entidades num texto. O método utilizado para realizar este processo é composto por um texto e alguns parâmetros de configuração, para que posteriormente possa ser ajustado às necessidades do utilizador. Para o processo implementado foram utilizados os valores do Quadro 4.8.

Quadro 4.8 - Parâmetros enviados em cada chamada do webservice do DS

Parâmetro	Valor
Text	Título + Corpo da Notícia
Confidence	0.2
Support	10

O parâmetro "**Confidence**" é o grau de certeza da anotação proposta pelo DS e o parâmetro "**Support**" é o número de ligações mínimas que a entidade na Dbpedia terá de ter para que seja considerado na análise.

4.3.2. Indexação de conteúdos a áreas

Após a finalização da fase 1, o próximo passo será selecionar todos os artigos no estado descrito no ponto 4.3.1 e indexar os que são relevantes às áreas indicadas no ponto 4.1.

Para aplicar o processo de indexação foi criado um serviço Windows que é executado num intervalo de tempo definido nas suas configurações. O serviço inicia o processo de indexação com a obtenção de todos os artigos que estejam no estado descrito no ponto 4.3.1 (passo 1 da Figura 4.7). A tarefa seguinte é obter todas as áreas e as respetivas definições armazenadas na tabela *AreaSearchProfile* (passo 2 da Figura 4.7). Para cada artigo obtido, o serviço aplica a definição de cada área sob a forma de um filtro no índice SOLR e no caso de existir uma correspondência positiva, o artigo é indexado à área (passo 3 da Figura 4.7).

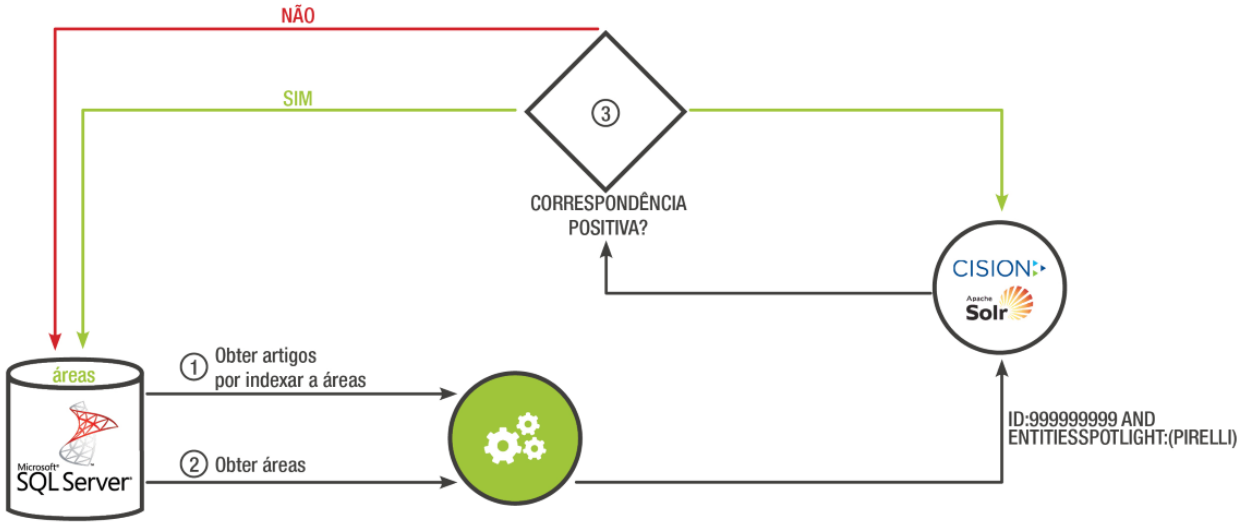


Figura 4.7 – Fase 2 do Processo de indexação com base no reconhecimento de entidades

Utilizando o artigo do SuperMotores.net incluído no ponto 4.3.1, o serviço identifica que o artigo se encontra disponível para indexação no passo 1 da Figura 4.7, de seguida obtém as seis áreas em análise e o artigo é validado, ocorrendo ambos no passo 2 da Figura 4.7. O processo repete-se para cada uma das áreas selecionadas no ponto 4.1, obtendo-se os resultados descritos no Quadro 4.9.

Quadro 4.9 – Resultado da indexação de um artigo para as áreas em estudo

Área	Filtro	Resultado
TAP	Id:48778768 AND EntitiesSpotlight:"TAP_Portugal"	Negativo
PT	Id:48778768 AND EntitiesSpotlight:"Portugal_Telecom"	Negativo
Pirelli	Id:48778768 AND EntitiesSpotlight:Pirelli	Positivo
BdP	Id:48778768 AND EntitiesSpotlight:"Banco_de_Portugal	Negativo
Air France	Id:48778768 AND EntitiesSpotlight:"Air_France"	Negativo
CR7	Id:48778768 AND EntitiesSpotlight:"Cristiano_Ronaldo"	Negativo

Analisando o quadro anterior, o serviço obteve uma resposta positiva para área Pirelli, procedendo à sua indexação.

4.3.3. Resultados

Após implementação das duas fases do processo descritas anteriormente nos pontos 4.3.1 e 4.3.2, procedeu-se à realização de um teste para aferir os resultados do sistema proposto para as áreas indicadas no ponto 4.1. As métricas utilizadas para verificar a qualidade do sistema foram a precision e o recall.

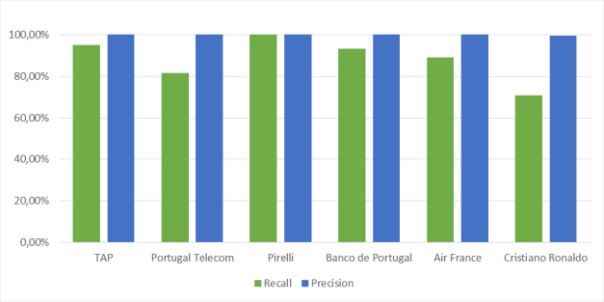


Gráfico 4.2 - Resultados no intervalo 05-08 a 12-08

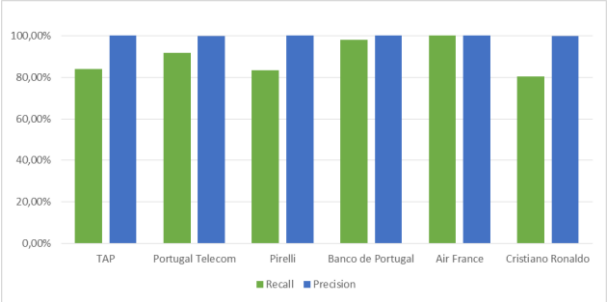


Gráfico 4.1 – Resultados no intervalo 13-08 a 20-08

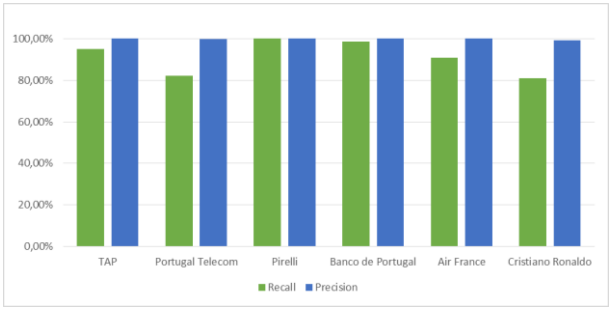


Gráfico 4.3 – Resultados no intervalo23-09 a 30-09

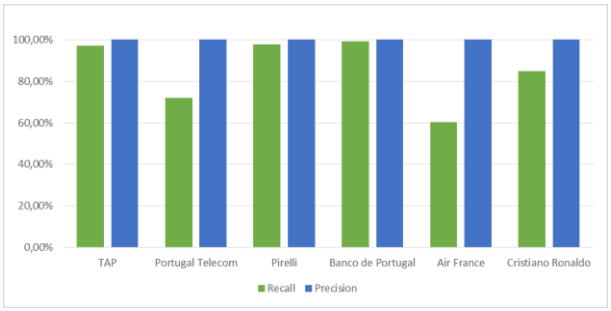


Gráfico 4.4 - Resultados no intervalo 07-10 a 14-10

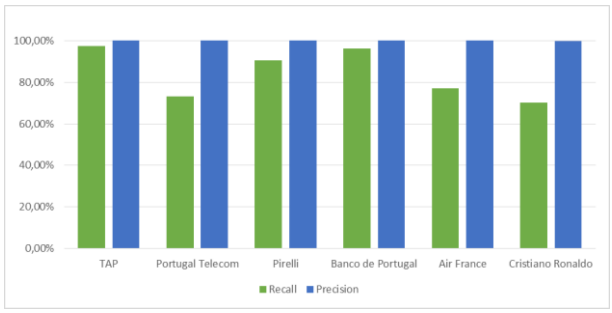


Gráfico 4.6 – Resultados no intervalo 18-11 a 25-11

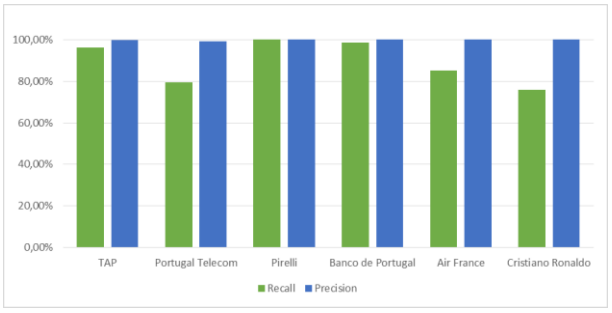


Gráfico 4.5 – Resultados no intervalo 09-12 a 16-12

Analisando os gráficos do Gráfico 4.2 Gráfico 4.5, verifica-se que métrica *precision* da DS para as seis áreas, nos seis intervalos de tempo, tem um valor muito próximo dos 100%. Para o caso de se fazer a média de todos os valores, obtemos uma *precison* de 99,90%, ou seja, o sistema falha uma proposta em cada mil. Quando se compara o número de artigos devolvidos por este método com os da produção, verificamos que existe alguma diferença, em média devolve menos 12%, significando um *recall* médio de 88%. Para compreender esta diferença procedeu-se à análise das falhas e foram identificadas as seguintes causas:

1. Erros sintáticos:
    a. Não existia um espaço entre um ponto final e a palavra inicial da frase seguinte;
    b. A entidade estava com a primeira letra em minúscula.
2. Definição incompleta da área:
    a. A área Air France tinha apenas “Air_France” como entidade da DbPedia na definição, mas num dos períodos de tempo, 40% das notícias publicadas eram sobre a empresa mãe Air France–KLM, correspondendo à entidade Air_France-KLM na DbPedia.

3. Falha na identificação da entidade:

    a. A entidade era reconhecida no texto, mas o algoritmo de desambiguação falha, indicando uma diferente. Por exemplo, na área Cristiano Ronaldo, em artigos onde fosse identificado a surfaceform "Ronaldo", a DS associava ao jogador brasileiro Ronaldo;

    b. O processo de identificação de entidades no texto falhava. O nome estava corretamente escrito e não existia qualquer erro sintático.

O processo implementado tem uma taxa de erro muito baixa nas indexações que faz, mas comparando com os números do sistema atual, falha doze artigos em cada cem. Após a análise das falhas conclui-se que umas poderiam ser solucionadas adicionando mais entidades na definição da área, por exemplo, no caso da área Air France, de acordo com o Quadro 4.6, a definição inicial tinha apenas a entidade Air France, mas no caso de se adicionar a entidade Air_France-KLM e combinar as duas entidades com o operador booleano *OR*, iria-se aumentar o *recall* sem prejudicar a *precision*.

Quadro 4.10 – Redefinição da área Air France

Área	Definição V1	Definição V2
Air France	Air_France	Air_France OR Air_France-KLM

A solução apresentada no Quadro 4.10 apenas solucionaria uma parte do problema, pois as restantes falhas continuariam a surgir.

Um dos objetivos iniciais da investigação seria evitar que as indexações fossem realizadas por meio de combinações de palavras-chave com operadores booleanos. Por forma a respeitar este pressuposto, adicionou-se uma nova fase no processo, caso o artigo seja descartado para a área em análise, mas ao mesmo tempo aumentar o *recall*:

1. Verificar a existência de referências às entidades que definem a área no texto;
2. Extrair as frases com as referências, caso a condição referida no ponto 1 seja verdadeira;
3. Averiguar qual o grau de similaridade com as outras frases dos artigos já indexados à referida área.

4.4. Sistema – Fase 2

Com base nas conclusões obtidas no ponto 4.3.3, é necessário criar um índice para aferir o grau de similaridade entre as frases dos artigos já indexados e as novas propostas. Procedeu-se à seleção de todos os artigos por área já indexados, à extração, remoção de stop-words e *stemização* das frases com referências às entidades incluídas na definição. Após a aplicação do pré-processamento, cada frase fica reduzida a uma lista de *tokens* que serão indexados a um índice do tipo *Latent Semantic Indexing* (LSI) através da utilização da biblioteca Gensim em

linguagem Python. Será criado um índice por área e esta nova fase do processo será denominada de WikiSim.

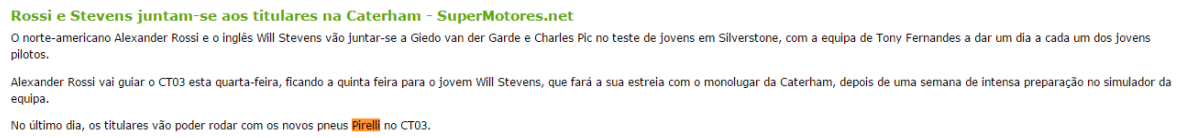
A integração do **WikiSim** no processo de indexação, descrito no ponto 4.3.2, foi feita através do desenvolvimento de um serviço REST em Python, sendo acedido por HTTP, através de pedidos POST por parte do serviço Windows.

Considerando a área Pirelli, o sistema indexou um total de 201 artigos (Quadro 4.11) no ponto 4.3.

Quadro 4.11 – Número de artigos extraídos para a área Pirelli em cada intervalo de tempo

05.08 A 12.08	**13.08 A 20.08**	**23.09 A 30.09**	**07.10 A 14.10**	**18.11 A 25.11**	**09.12 A 16.12**
18	20	17	44	58	44

Dos artigos indexados extraíram-se as frases que incluíssem a expressão de texto "Pirelli" no título ou no corpo da notícia. No total identificaram-se 397 frases e uma a uma foram transformadas numa lista de tokens através de remoção das stop-words, alteração das palavras para letras minúsculas e as derivadas para a sua forma base. O resultado de cada frase é indexado ao índice.


Rossi e Stevens juntam-se aos titulares na Caterham - SuperMotores.net

O norte-americano Alexander Rossi e o inglês Will Stevens vão juntar-se a Giedo van der Garde e Charles Pic no teste de jovens em Silverstone, com a equipa de Tony Fernandes a dar um dia a cada um dos jovens pilotos.

Alexander Rossi vai guiar o CT03 esta quarta-feira, ficando a quinta feira para o jovem Will Stevens, que fará a sua estreia com o monolugar da Caterham, depois de uma semana de intensa preparação no simulador da equipa.

No último dia, os titulares vão poder rodar com os novos pneus Pirelli no CT03.

Figura 4.8 – Identificação no artigo SuperMotores.net da frase com a expressão Pirelli

Para a demonstração do processo de indexação, utilizou-se o artigo do ponto 4.3.1. A frase identificada na Figura 4.8 que incluía a expressão de texto "Pirelli" foi extraída. Após a aplicação do processo descrito, a frase "***No último dia, os titulares vão poder rodar com os novos pneus Pirelli no CT03.***" é transformada na seguinte lista de tokens: "**últim titul rod pneus pirell ct03**".

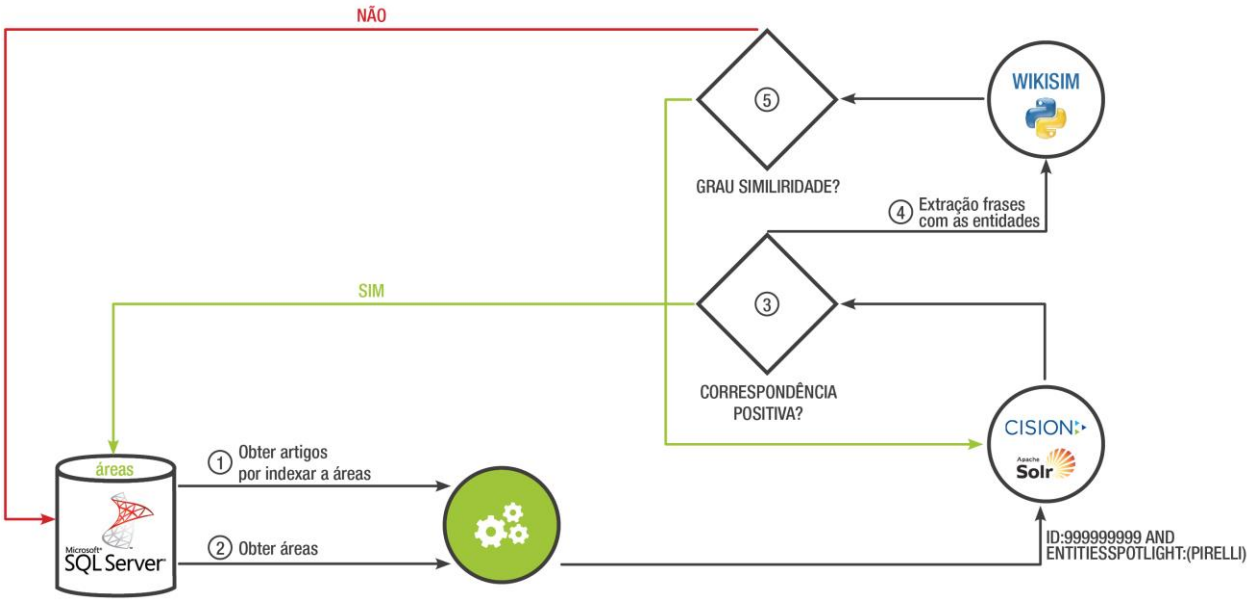


Figura 4.9 – Fase 2 do processo atualizado com adição do serviço o WikiSim

No caso do artigo não ser indexado à área em análise no passo 3 da Figura 4.9, mas incluir as entidades no texto, procede-se à extração, agrupamento e envio das frases em forma de lista de tokens para o serviço WikiSim. As frases recebidas são comparadas uma a uma com o índice, de forma a obter o grau de similaridade com todas as frases incluídas no mesmo. O método que estabelece a comparação devolve dois valores:

- $\overline{x}_{simscoref}$: a média aritmética dos graus similaridade que a frase tem com cada elemento do índice (11); (11)

$$\overline{x}_{simscoref} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

- $\overline{x}_{top10scoref}$: os graus de similaridade são ordenados de forma decrescente e extraindo-se uma amostra de 10% do número total de itens do índice e é calculado uma média aritmética (12);

$$\overline{x}_{top10scoref} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \quad (12)$$

Na (11) e (12) o n representa o número total de documentos que compõe o índice de similaridade e o x_n é o grau de similaridade entre o documento n e a lista de tokens em análise.

Após a obtenção dos scores $\overline{x}_{simscoref}$ e $\overline{x}_{top10scoref}$,para cada lista de *tokens*, o WikiSim calcula a média aritmética de todas as listas que compõe o documento, por forma a obter um *score* global do documento.

$$\overline{x}_{simdoc} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \quad (14)$$

$$\overline{x}_{simtop10doc} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \quad (13)$$

Na (13) e (14) o n representa o número de listas de *tokens* enviadas pelo serviço Windows e o x_n é o grau de similaridade entre cada lista e o índice documentos.

O serviço Windows após obtenção dos dois scores verifica se score $\overline{x}_{simtop10doc}$ é maior ou igual ao valor mínimo, definido previamente para cada área, no caso de ser verdadeiro, o documento é indexado à área, caso contrário, é calculado um score global de similaridade para o documento (15), através da média aritmética entre o score $\overline{x}_{simdoc}$ e o $\overline{x}_{simtop10doc}$. No caso do $\overline{x}_{scoreglobal}$ ser maior 0,056, o artigo é indexado à área.

$$\overline{x}_{scoreglobal} = \frac{1}{2}\left(\overline{x}_{simdoc} + \overline{x}_{simtop10doc}\right) \quad (15)$$

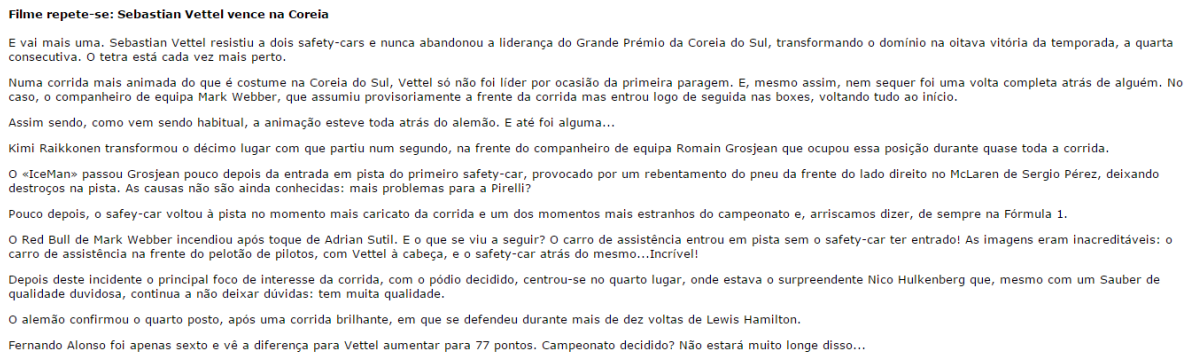
Os limites de aceitação dos scores $\overline{x}_{simdoc}$, do $\overline{x}_{simtop10doc}$ e do $\overline{x}_{scoreglobal}$ foram definidos após várias execuções dos testes, de forma a obter o melhor compromisso entre as métricas *recall* e *precsion*. Os scores $\overline{x}_{simdoc}$ e o $\overline{x}_{simtop10doc}$ são definidos por área (ver

Quadro 4.12), mas no caso do $\overline{x}_{scoreglobal}$ o valor mínimo é comum para todas as áreas.

Quadro 4.12 – $\overline{x}_{\mathrm{simtop10doc}}$ por área

Área	$\overline{x}_{simtop10doc}$
TAP	0,16
PT	0,31
Pirelli	0,30
BdP	0,10
Air France	0,20
CR7	0,17

Para exemplificação do processo, selecionou-se um artigo que não fosse indexado à área Pirelli pelo sistema no ponto 0. O artigo selecionado foi o "Filme repete-se: Sebastian Vettel vence na Coreia" do sítio Pi-racing.com (Figura 4.10).



Filme repete-se: Sebastian Vettel vence na Coreia

E vai mais uma. Sebastian Vettel resistiu a dois safety-cars e nunca abandonou a liderança do Grande Prémio da Coreia do Sul, transformando o domínio na oitava vitória da temporada, a quarta consecutiva. O tetra está cada vez mais perto.

Numa corrida mais animada do que é costume na Coreia do Sul, Vettel só não foi líder por ocasião da primeira paragem. E, mesmo assim, nem sequer foi uma volta completa atrás de alguém. No caso, o companheiro de equipa Mark Webber, que assumiu provisoriamente a frente da corrida mas entrou logo de seguida nas boxes, voltando tudo ao início.

Assim sendo, como vem sendo habitual, a animação esteve toda atrás do alemão. E até foi alguma...

Kimi Raikkonen transformou o décimo lugar com que partiu num segundo, na frente do companheiro de equipa Romain Grosjean que ocupou essa posição durante quase toda a corrida.

O «IceMan» passou Grosjean pouco depois da entrada em pista do primeiro safety-car, provocado por um rebentamento do pneu da frente do lado direito no McLaren de Sergio Pérez, deixando destroços na pista. As causas não são ainda conhecidas: mais problemas para a Pirelli?

Pouco depois, o safey-car voltou à pista no momento mais caricato da corrida e um dos momentos mais estranhos do campeonato e, arriscamos dizer, de sempre na Fórmula 1.

O Red Bull de Mark Webber incendiou após toque de Adrian Sutil. E o que se viu a seguir? O carro de assistência entrou em pista sem o safety-car ter entrado! As imagens eram inacreditáveis: o carro de assistência na frente do pelotão de pilotos, com Vettel à cabeça, e o safety-car atrás do mesmo...Incrível!

Depois deste incidente o principal foco de interesse da corrida, com o pódio decidido, centrou-se no quarto lugar, onde estava o surpreendente Nico Hulkenberg que, mesmo com um Sauber de qualidade duvidosa, continua a não deixar dúvidas: tem muita qualidade.

O alemão confirmou o quarto posto, após uma corrida brilhante, em que se defendeu durante mais de dez voltas de Lewis Hamilton.

Fernando Alonso foi apenas sexto e vê a diferença para Vettel aumentar para 77 pontos. Campeonato decidido? Não estará muito longe disso...

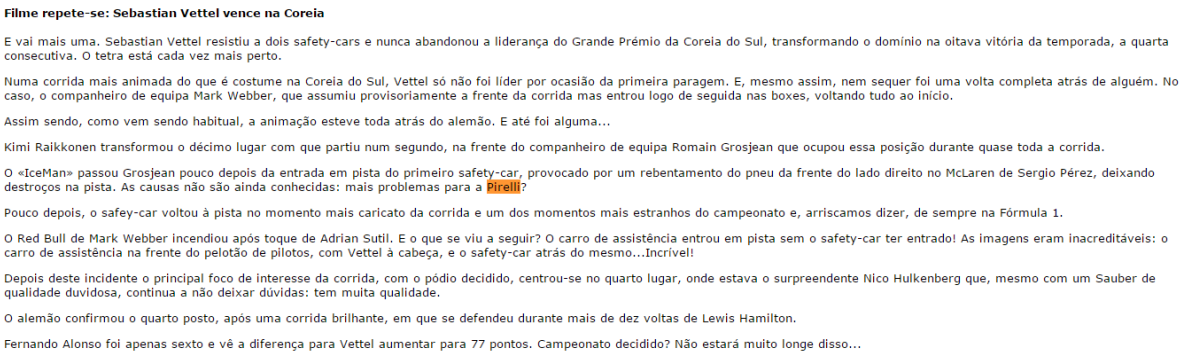
Figura 4.10 – Título e corpo da notícia do sítio Pi-racing.com

O artigo da Figura 4.10 falha indexação no passo 3 do processo da Figura 4.9, porque a entidade Pirelli não foi identificada pela DS no texto. Podemos verificar no quadro seguinte, a lista das entidades armazenadas no campo "EntitiesSpotlight" do índice do servidor SOLR e a não existência da entidade Pirelli.

Quadro 4.13 – Lista de entidades identificadas no artigo do sítio Pi-racing.com

Sebastian_Vettel	Adrian_Sutil	Nico_Rosberg
Mark_Webber	Pelotão_(ciclismo)	Jenson_Button
Filme	Nicolas_Hülkenberg	Felipe_Massa
Alemanha	Sauber	Williams_F1
Kimi_Räikkönen	Lewis_Hamilton	Pastor_Maldonado
Romain_Grosjean	Fernando_Alonso	Caterham_F1_Team
Homem_de_Gelo	Renault_F1	Van
Pneu	Mercedes-Benz	Jules_Bianchi
McLaren	Scuderia_Ferrari	Marussia_F1_Team
Scuderia_Toro_Rosso	Force_India	

O sistema verifica se o texto (título ou corpo da notícia) contém referências à entidade Pirelli, tal como mencionado no ponto 4.3.3 e de acordo com o passo 4 da Figura 4.9. No artigo utilizado para o exemplo, o sistema extrai uma frase conforme a Figura 4.11.



Filme repete-se: Sebastian Vettel vence na Coreia

E vai mais uma. Sebastian Vettel resistiu a dois safety-cars e nunca abandonou a liderança do Grande Prémio da Coreia do Sul, transformando o domínio na oitava vitória da temporada, a quarta consecutiva. O tetra está cada vez mais perto.

Numa corrida mais animada do que é costume na Coreia do Sul, Vettel só não foi líder por ocasião da primeira paragem. E, mesmo assim, nem sequer foi uma volta completa atrás de alguém. No caso, o companheiro de equipa Mark Webber, que assumiu provisoriamente a frente da corrida mas entrou logo de seguida nas boxes, voltando tudo ao início.

Assim sendo, como vem sendo habitual, a animação esteve toda atrás do alemão. E até foi alguma...

Kimi Raikkonen transformou o décimo lugar com que partiu num segundo, na frente do companheiro de equipa Romain Grosjean que ocupou essa posição durante quase toda a corrida.

O «IceMan» passou Grosjean pouco depois da entrada em pista do primeiro safety-car, provocado por um rebentamento do pneu da frente do lado direito no McLaren de Sergio Pérez, deixando destroços na pista. As causas não são ainda conhecidas: mais problemas para a Pirelli?

Pouco depois, o safey-car voltou à pista no momento mais caricato da corrida e um dos momentos mais estranhos do campeonato e, arriscamos dizer, de sempre na Fórmula 1.

O Red Bull de Mark Webber incendiou após toque de Adrian Sutil. E o que se viu a seguir? O carro de assistência entrou em pista sem o safety-car ter entrado! As imagens eram inacreditáveis: o carro de assistência na frente do pelotão de pilotos, com Vettel à cabeça, e o safety-car atrás do mesmo...Incrível!

Depois deste incidente o principal foco de interesse da corrida, com o pódio decidido, centrou-se no quarto lugar, onde estava o surpreendente Nico Hulkenberg que, mesmo com um Sauber de qualidade duvidosa, continua a não deixar dúvidas: tem muita qualidade.

O alemão confirmou o quarto posto, após uma corrida brilhante, em que se defendeu durante mais de dez voltas de Lewis Hamilton.

Fernando Alonso foi apenas sexto e vê a diferença para Vettel aumentar para 77 pontos. Campeonato decidido? Não estará muito longe disso...

Figura 4.11 – Identificação da frase que incluí a entidade Pirelli

A frase extraída é processada e transformada numa lista de *tokens*, o resultado final encontra-se descrito no Quadro 4.14, correspondendo ao passo 4 da Figura 4.9.

Quadro 4.14 – Lista de tokens extraídos do artigo do sítio Pi-racing.com

#	Frase	Tokens
1	As causas não são ainda conhecidas: mais problemas para a Pirelli?	caus conhec problem pirell

A lista de *tokens* criada é enviada para o WikiSim. O serviço, ao receber a lista compara-a com os 397 documentos que compõe o índice e no fim da comparação obtém uma lista com os 397 graus similaridade ordenados descendentemente.

Para o cálculo da $\overline{x}_{simscoref}$ faz-se a média aritmética dos 397 scores. Para obter o $\overline{x}_{top10scoref}$, o sistema extrai os 40 graus mais similares (10% do total de documentos que compõe o índice) e faz a sua média aritmética (15). Tendo-se obtido os valores do Quadro 4.15.

Quadro 4.15 – Score de similaridade da lista um de *tokens* para com o índice similaridade

#	$\overline{x}_{simscoref}$	$\overline{x}_{top10scoref}$
1	0.18	0.36

O processo era repetido por cada lista de *tokens* enviada para o WikiSim, como o exemplo tinha apenas uma lista, o processo inicia a aplicação das equações (14) e (13), para o cálculo dos *scores* $\overline{x}_{simdoc}$ e do $\overline{x}_{simtop10}$. Utilizando os valores do Quadro 4.15 obtiveram-se os seguintes resultados:

$$\overline{x}_{simdoc} = \frac{1}{1}(0{,}18) = 0{,}18$$

$$\overline{x}_{simtop10} = \frac{1}{1}\sum_{i=1}^{1} x_i = \frac{1}{1}(0{,}36) = 0{,}36$$

Para que o artigo possa ser indexado à área Pirelli, é necessário que uma das condições seja verdadeira:

1. $\overline{x}_{simtop10} \geq 0{,}30$
2. $\overline{x}_{scoreglobal} \geq 0{,}056$

Verifica-se que a primeira condição é verdadeira, procedendo-se à indexação do artigo à área. No caso da condição 1 ser falsa, procedia-se à aplicação da (15) para obtenção do *score* global de similaridade do documento.

$$\overline{x}_{scoreglobal} = \frac{1}{2}(0{,}18 + 0{,}36) = 0{,}27$$

O sistema tomaria também a decisão de indexar o artigo à área Pirelli porque o $\overline{x}_{scoreglobal} \geq 0{,}056$.

4.4.1. Resultados

Após implementação do processo descrito no ponto 4.4, procedeu-se à realização de um novo teste para verificar se o *recall* obtido com a primeira versão do sistema era melhorado, mas sem prejudicar em demasia a *precision* original.

De acordo com os gráficos Gráfico 4.8 aoGráfico 4.11, com a implementação deste passo no sistema consegue-se aumentar o *recall* médio de 88% para 98,89%, com uma diminuição aceitável da *precision* média em 1,03%.

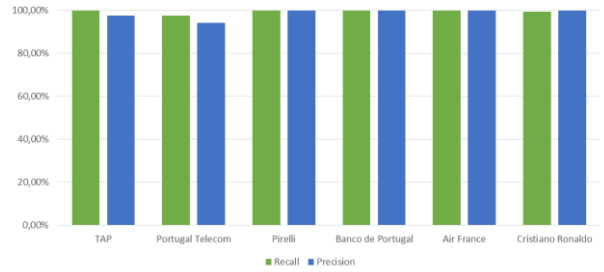


Gráfico 4.8 – Resultados no intervalo 05-08 a 12-08

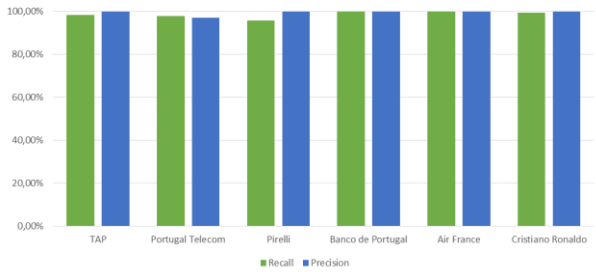


Gráfico 4.7 – Resultados no intervalo 13-08 a 20-08

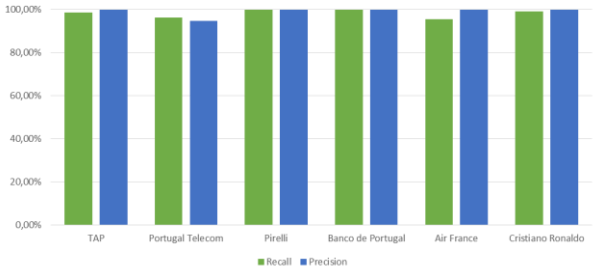


Gráfico 4.10 - Resultados no intervalo 23-09 a 30-09

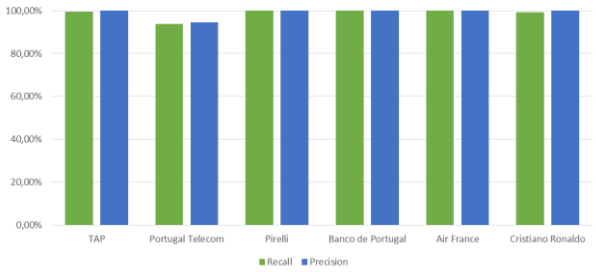


Gráfico 4.9 – Resultados no 07-10 a 14-10

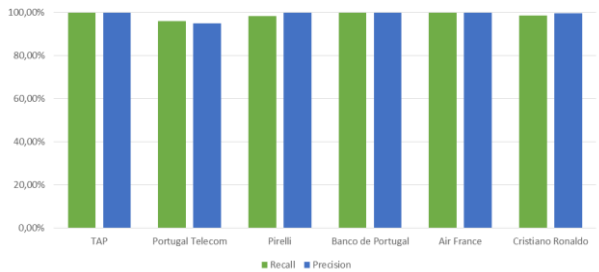


Gráfico 4.12 – Resultados no intervalo 18-11 a 25-11

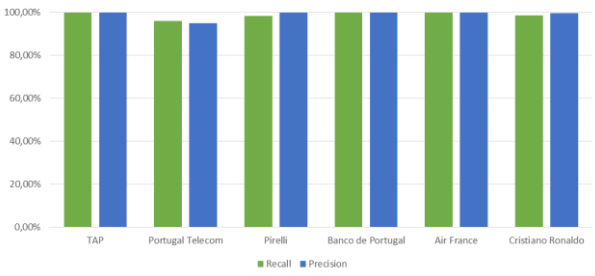


Gráfico 4.11 – Resultados no intervalo 09-12 a 16-12

No seguimento das conclusões nas áreas automáticas selecionadas para a investigação, decidiu se validar este processo para uma área que necessita de validação por parte das equipas de produção devido à sua ambiguidade. A área selecionada foi o Continente, que agrega todos os artigos relacionados com os Hipermercados Continente.

4.5. Área Continente

De acordo com as conclusões obtidas no ponto 4.3.3, o primeiro passo foi a validação do recall e a da precison do sistema DS para a nova área. O sistema obteve uma média de 12,44% de recall e 93.15% de precision. Como os resultados obtidos são insatisfatórios (Gráfico 4.13), procedeu-se à análise dos artigos aceites para identificar a expressão de texto (*surface form*) que permitiu a indexação à área Continente. Verificou-se que 64% dos artigos indexados continham a surface form "Missão Sorriso" e os restantes 36% pela surface form "hipermercado Continente", mas no caso do artigo conter a *surface form* "hipermercados Continente", o sistema falha a associação.

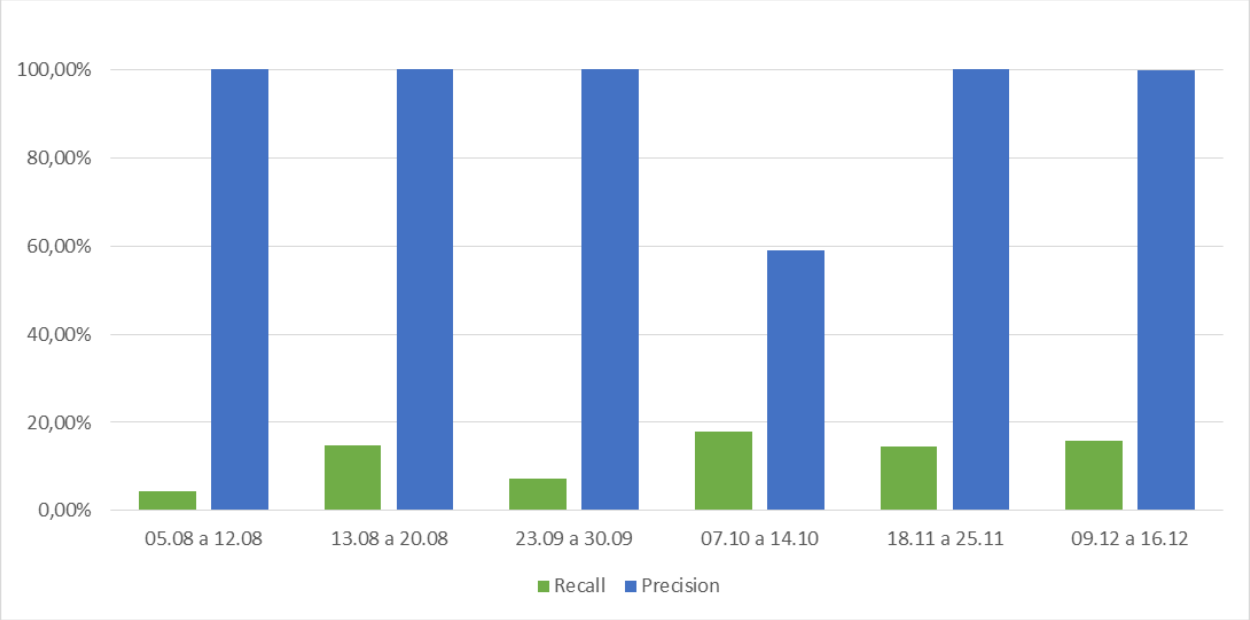


Gráfico 4.13 - Resultados da área Continente nos 6 períodos de tempo

O passo seguinte seria aplicar a versão 2 do sistema, contudo o número de artigos identificados pelo DS, para a área Continente, foi em número insuficiente para criar um índice de similaridade suficientemente bom para distinguir/desambiguar os artigos corretamente. Surge a dúvida de como se poderia obter um conjunto de artigos que permitissem fazer a desambiguação da área Continente e que respeitasse as premissas iniciais da investigação. A solução encontrada foi aplicar um pouco do conceito usado no ponto 4.4, mas com uma variação: um artigo só seria considerado válido, se no texto da notícia (título e corpo da notícia) contivesse a palavra "Continente" com a letra C em maiúscula e que o sistema DS tivesse identificado entidades que fizessem parte da definição da entidade Hipermercados Continente na Dbpedia.

O primeiro passo foi obter a lista das entidades associadas à entidade Hipermercados Continente. Optou-se por usar a API disponibilizada pela Wikipedia, devido ao menor grau de complexidade que se exige para uma futura integração num sistema de produção. Apesar da DS usar os recursos da Dbpedia para fazer a anotação e desambiguação, as definições usadas por esta foram extraídas da Wikipedia. Os nomes das entidades na DbPedia são exatamente iguais aos da Wikipedia. A API retornou quarenta e nove entidades, mas apenas quatro foram selecionadas:

1. Sonae
2. Continente Modelo
3. Sonae Distribuição
4. Sonae MC

No processo de seleção das entidades mais relevantes foram tidos em conta os seguintes fatores:

1. Nome das entidades representar um grau de ambiguidade baixa;
2. Forte relacionamento com os Hipermercados Continente.

Outra restrição adicionada ao filtro na seleção do corpus, para aplicação na desambiguação, foi a remoção de todos os artigos em que o DS tivesse identificado a entidade "Continente" nos seus textos, na tentativa de evitar que sejam incluídos artigos que iriam gerar erros no processo de desambiguação.

Aplicado o filtro descrito nos seis períodos de tempo e adicionados os artigos previamente identificados pela DS como sendo Hipermercados Continente, obteve-se um corpus de 311 artigos.

Nos testes aplicados no primeiro intervalo de tempo obteve-se um *recall* muito próximo dos 98% e uma *precision* de 70%, como se observa no Gráfico 4.14. O passo seguinte foi validar o sistema para segundo e terceiro intervalos de tempo e o comportamento foi idêntico ao do primeiro. Um *recall* muito próximo dos 100%, mas uma baixa *precision*. Ao verificar-se este comportamento, decidiu-se não realizar os testes nos restantes intervalos. O sistema, nos 3 primeiros intervalos, obteve em média um *recall* de 98,83% e uma *precision* média de 66,60%.

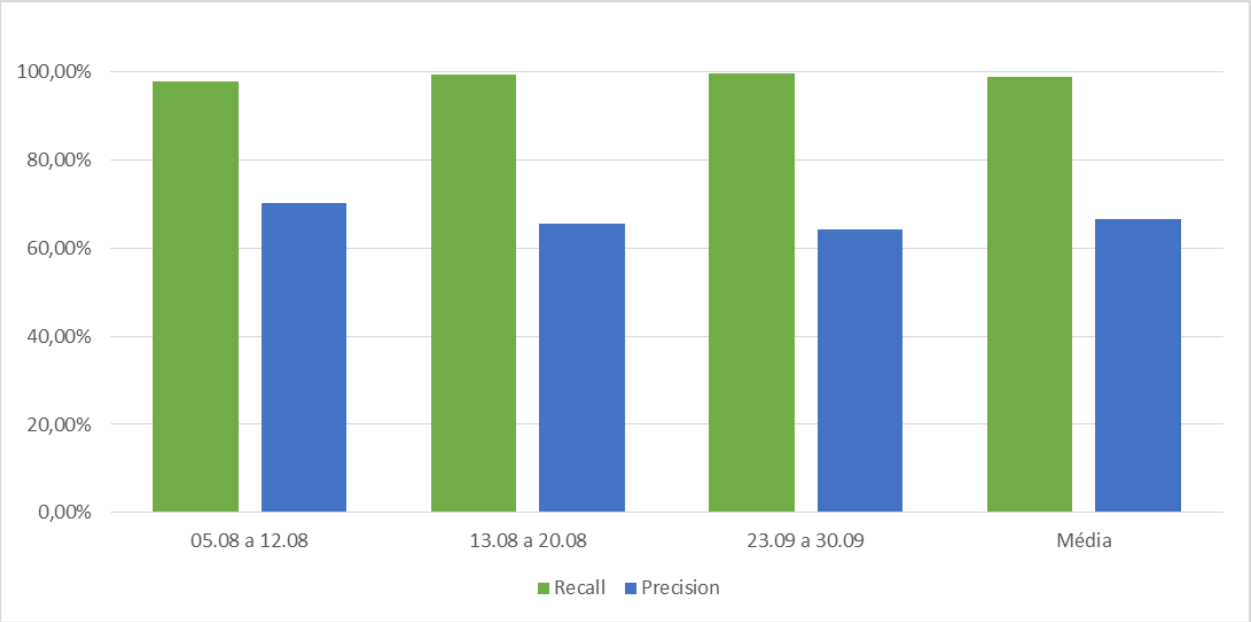


Gráfico 4.14 - Resultados da área Continente nos três primeiros períodos de tempo

Para identificar as causas para o elevado número de falsos positivos, verificou-se que no caso dos artigos que contivessem a expressão de texto "Continente" e o âmbito se referisse a assuntos relativos a Portugal Continental (meteorologia, fogos, variação de preços e etc), o processo de desambiguação falhava.

Em suma, o sistema proposto para as áreas automáticas, não poderá ser aplicado nesta área porque têm um elevando grau de ambiguidade. Desta forma surge interesse por parte da CISION em prosseguir a investigação para avaliar outros métodos para áreas idênticas à do Continente.

4.5.1. Sistema – Fase 3

Para solucionar o problema descrito no ponto 4.4.1 para a área Continente, implementou-se dois métodos:

1. Criar um índice de similaridade de textos completos, em vez de um com frases;
2. Classificação automática de documentos.

Índice de similaridade de textos completos

Para o desenvolvimento do índice similaridade de textos completos para a área Continente, utilizou-se o corpus de 311 artigos obtidos através do método descrito no ponto 4.5. Ao corpus aplicou-se o mesmo pré-processamento utilizado na implementação do índice similaridade por frases no ponto 4.4: remoção das stop-words, stemização, transformação de cada documento numa lista de tokens e a sua indexação ao índice do tipo LSI.

Após a implementação, procedeu-se à validação do novo índice e verificou-se um comportamento oposto ao do índice de similaridade de frases. Obteve-se uma *precision* média de 94% e um *recall* médio de 56% (Gráfico 4.15). Através dos resultados confirmou-se que este tipo de abordagem não poderá ser aplicado na área Continente, devido ao seu elevado grau de ambiguidade.

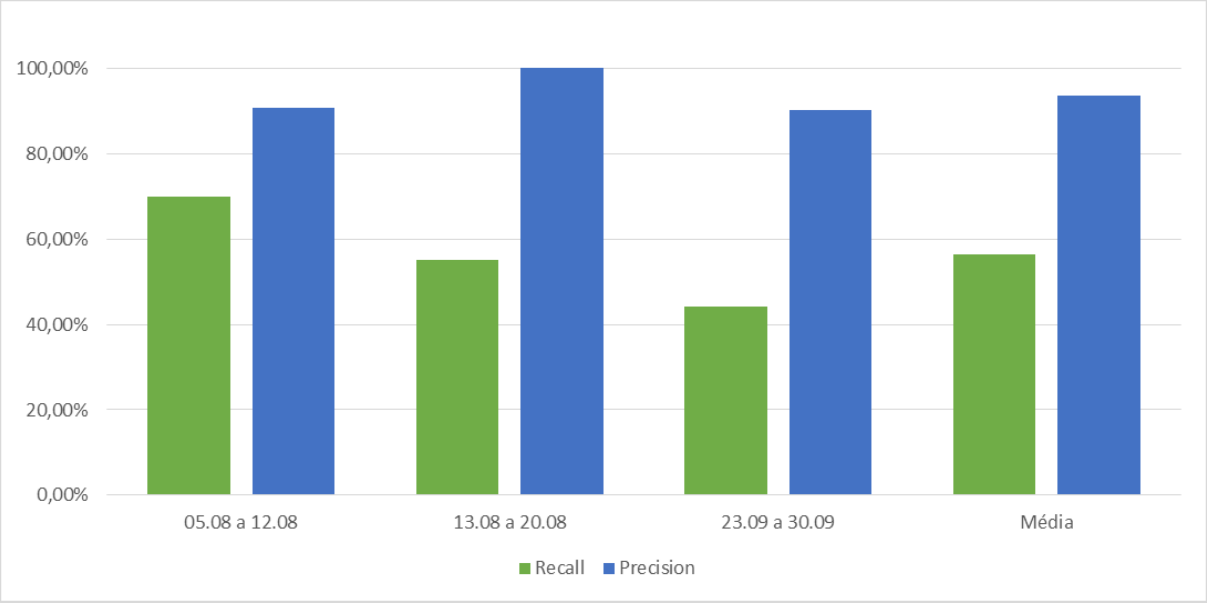


Gráfico 4.15 - Resultados da área Continente nos três primeiros períodos de tempo, para um índice de similaridade com o texto completo

Classificação Automática de Textos

Seguiu-se o passo de implementação de um sistema de classificação automática de textos para realizar a desambiguação da entidade Hipermercados Continente. Para o desenvolvimento deste processo utilizou-se a biblioteca NLTK em Python. Segundo (Perkins, 2010) uma forma de melhorar a performance de um sistema de classificação automática de textos é combinar vários algoritmos através do método de votação, onde a classe vencedora será aquela que obter mais votos. Neste tipo de abordagem, o número de algoritmos selecionados deverá ser impar, por forma a evitar empates.

O autor (Perkins, 2010) propõe que o sistema inclua os seguintes classificadores:

1. Naive Bayes;
2. Decision Tree;
3. Maxent.

O classificador por votação que agrega os três classificadores vai ser identificado no presente documento como MaxVote (Perkins, 2010).

Os três classificadores para serem desenvolvidos necessitam de ser treinados com um corpus, onde cada texto está já classificado com a classe correta. Para o caso em investigação foi necessário criar um corpus com duas classes: Hipermercados Continente e Geografia. A classe Hipermercados Continente inclui os artigos que são relevantes para o mesmo. Os artigos que classificam a classe Geografia devem ser dos restantes significados da palavra Continente. Os classificadores, neste tipo de ambiente, onde apenas se identifica duas classes são qualificados como binários.

Os artigos utilizados para a classe "Hipermercados Continente" foram os selecionados no ponto 4.5 para implementação de um índice de similaridade, tendo-se obtido um total de 311 artigos.

As condições definidas para a seleção de artigos da classe "Geografia" são:

- Não incluir nenhuma das cinco entidades que fazem parte da definição da entidade "Hipermercados Continente" no ponto 4.5;
- DS ter identificado no corpo da notícia pelo menos uma das seguintes entidades:
    - Continente;
    - Ásia;
    - Europa;
    - América do Norte;
    - América do Sul;
    - América Central;
    - Oceânia;
    - Antártida;
    - África.

As condições para a seleção dos artigos da classe Geografia foram aplicadas aos seis intervalos de tempo em estudo e obteve-se de uma forma aleatória 300 artigos. O corpus para treinar os classificadores automáticos de texto foi criado com um total de 611 artigos.

Finalizado o processo de elaboração do corpus, procedeu-se a um pré-processamento dos textos, aplicando-se as seguintes tarefas:

1. Segmentação do texto;

2. Remoção de palavras não determinantes (Stop Words);
3. Seleção de Características (bag of words).

A primeira tarefa, segmentação do texto, tem como objetivo dividir o texto em palavras através da localização dos limites entre estas. No caso da língua portuguesa é o espaço e os sinais de pontuação. Após a obtenção da lista de palavras por classe que compõe o corpus, procedeu-se à remoção de todas as palavras não descriminantes (stop words), ou seja, aquelas que não adicionam qualquer valor à análise. Para finalizar o pré-processamento do corpus, na seleção das características mais relevantes, o autor (Perkins, 2010) propõe um método que realize a contagem da frequência de cada palavra, bem como a frequência condicional de cada palavra dentro de cada classe. Para o cálculo da pontuação de cada palavra, utilizou-se a métrica Chi-quadrado com os seguintes parâmetros:

- n_ii: frequência da palavra na classe;
- n_ix: frequência total no corpus;
- n_xi: frequência total de todas as palavras na classe;
- n_xx: frequência total de todas as palavras no corpus.

A maneira mais simples de pensar sobre estes parâmetros é que quanto mais perto estiver o n_ii do n_ix, maior é a pontuação, ou seja, quanto mais frequente é uma palavra numa classe, relativamente à sua ocorrência no corpus, maior é a sua pontuação. Assim que se obtiver as pontuações de cada palavra em cada classe, pode-se filtrar todas as palavras cuja pontuação é inferior ao limite definido. Para implementação dos classificadores considerou-se o valor de cinco.

O passo seguinte foi o desenvolvimento dos modelos para os classificadores com base no resultado obtido do processo de pré-processamento. Estes foram criados tendo em conta as seguintes características:

- Correção de *Laplace* foi usada no classificador *Naive Bayes* para o caso de uma das probabilidades ser nula;
- Classificador *Decision Tree* foi criado utilizando os valores por defeito do classificador na biblioteca NLTK;
- Classificador *Maxent* foi criado utilizando o algoritmo *GIS - Generalized Iterative Scaling* e com cem interações no máximo;
- Utilização de 75% e 25% do conjunto de treino para criar e testar os modelos respetivamente.

Para a avalidação dos classificadores foram usadas as seguintes métricas: *accuracy*, *precision* e *recall*. Os valores obtidos estão descritos no Quadro 4.16.

Quadro 4.16 – Métricas obtidas para cada classificador do sistema de votação e as métricas

Algoritmo	Accuracy	Precision		Recall	
		Geografia	Hipermercados Continente	Geografia	Hipermercados Continente
Naive Bayes	0,95	0,915	1	1	0,91
Decsion Tree	1	1	1	1	1
Maxent	0,94	0,914	0,986	0,986	0,91
MaxVote	0,96	0,926	1	1	0,923

O classificador MaxVote foi incluído no serviço WikiSim, de forma a que o serviço Windows consiga chamar o classificador e decidir se associa ou não os artigos à área Hipermercados Continente. O sistema ficou de acordo com a arquitetura da Figura 4.12.

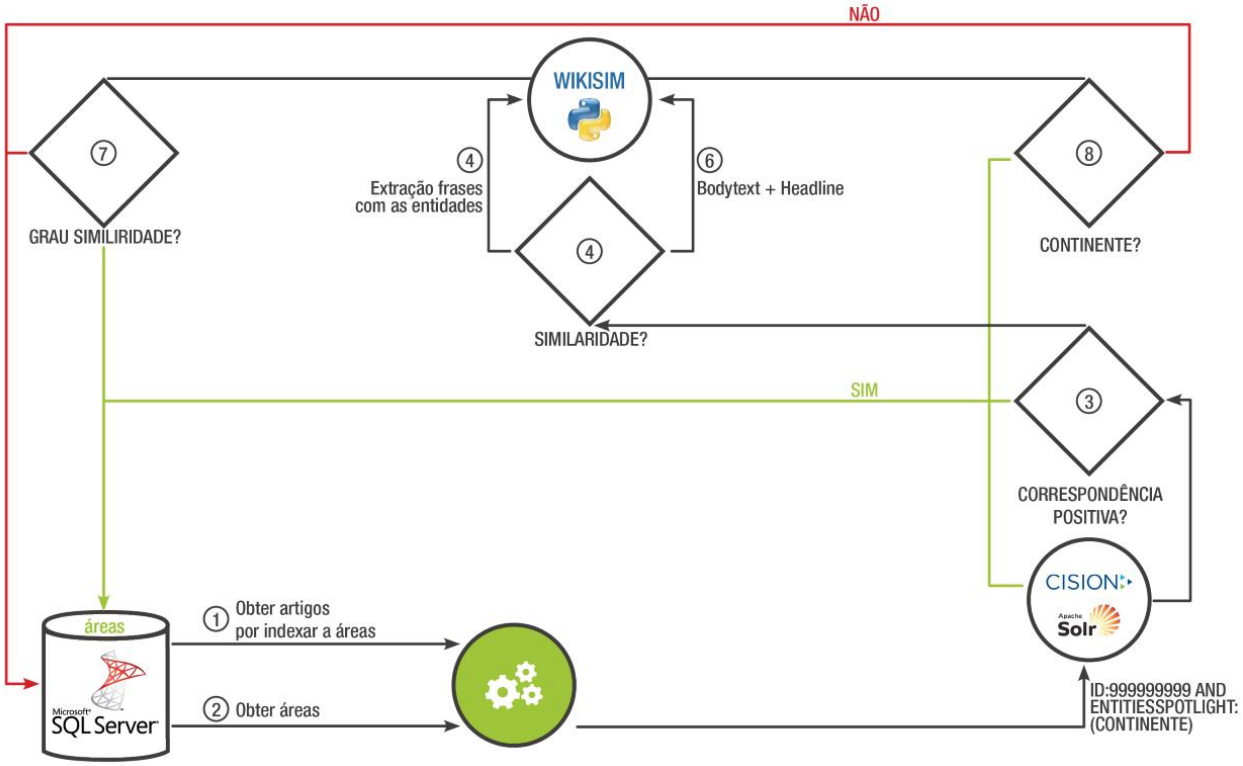


Figura 4.12 - Fase 3 do processo atualizado com decisão entre dois métodos de desambiguação do serviço WikiSim

O método de validação será definido por área. Procedeu-se à validação da nova arquitetura do sistema e os resultados estão descritos no Gráfico 4.16.

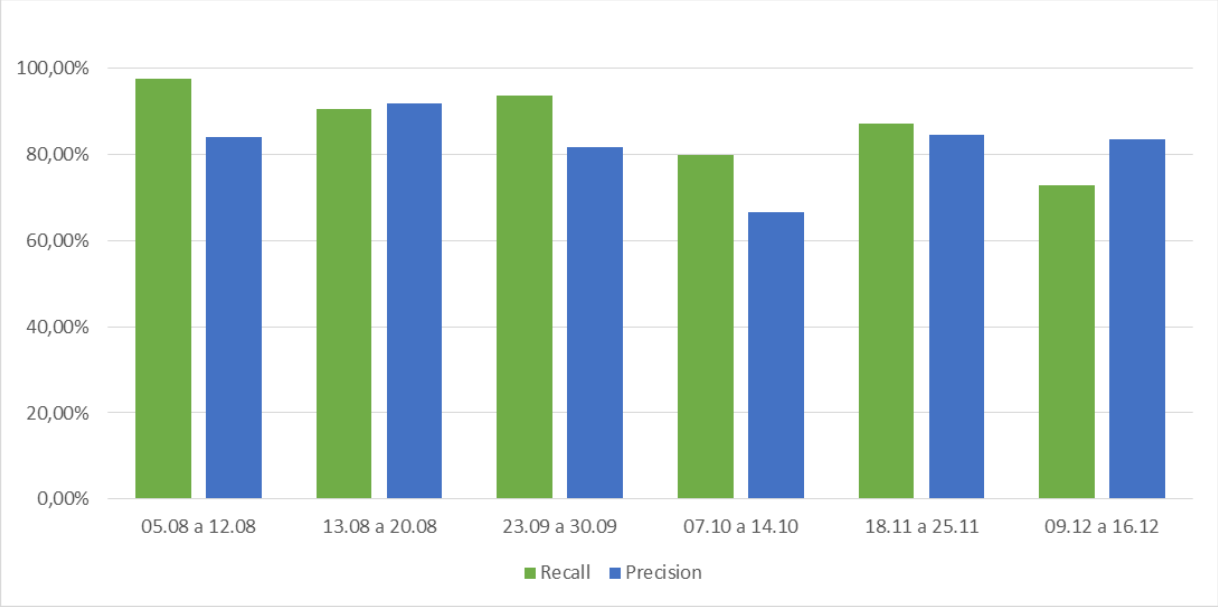


Gráfico 4.16 - Resultados da área Continente nos seis períodos de tempo para a fase 3 do sistema

Comparando as três abordagens nos três primeiros intervalos, descritos no Quadro 4.17, verifica-se que o classificador MaxVote é o método mais equilibrado, obtendo um *recall* médio 93,78% e uma *precision* de 85,76%. Comparando o método de similaridade de frases com o MaxVote conseguiu-se um aumento próximo dos 20% na métrica *precision* e com uma perda de apenas 5% no *recall*.

Quadro 4.17 – Comparação das métricas entre os três métodos usados para a desambiguação da área Hipermercados Continente

	05.08 A 12.08		**13.08 A 20.08**		**23.09 A 30.09**		**Média**	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Similaridade através de frases	97,65%	70,03%	99,35%	70,03%	99,49%	64,26%	98,83%	66,40%
Similaridade através do texto completo	69,84%	100%	55,00%	100%	44,05%	90,24%	56,30%	93,66%
Classificador MaxVote	97,35%	84,35%	90,41%	91,60%	93,58%	81,60%	93,78%	85,76%

Caso a análise abranja os seis períodos de tempo, verifica-se que o MaxVote continua a garantir um compromisso aceitável entre as duas métricas. Obtendo um valor médio do *recall* de 86,82% e uma *precision* média de 81,94%. Contudo a taxa de erro é um pouco elevada, na ordem dos

18%, então na tentativa de efetuar um melhoramento das métricas, decide-se implementar um novo classificador Naive Bayes, mas modificando o pré-processamento do corpus:

1. Segmentação do texto;
2. Remoção de palavras não determinantes (Stop Words);
3. Normalização das palavras:
    a. Conversão para minúsculas;
    b. Stemização.
4. Seleção de Características (bag of words).

Na implementação do pré-processamento adicionou-se a normalização das palavras do corpus e alterou-se a métrica de seleção de características. Na tarefa de normalização das palavras, estas foram transformadas em minúsculas e as derivadas na sua forma base, através da aplicação da técnica *stemming*. Na seleção de características foi aplicada a métrica frequência das palavras no texto. Para a validação do classificador foram usadas as seguintes métricas: accuracy, precision e recall. Os valores obtidos encontram-se enumerados no Quadro 4.18.

Quadro 4.18 – Métricas obtidas para o classificador

Accuracy	**Precision**		**Recall**	
	Geografia	**Hipermercados Continente**	**Geografia**	**Hipermercados Continente**
0,91	0,95	0,89	0,85	0,97

O classificador foi adicionado ao serviço WikiSim e o serviço Windows para associar um artigo à área Continente, terá de obter uma votação unânime. Verificou-se que existiu um aumento na métrica *precision* na ordem dos 4,7%, mas em contrapartida perdeu-se no *recall* aproximadamente 6,1% dos artigos.

Numa última tentativa para obter um melhor compromisso entre as duas métricas adicionou-se um novo classificador com base na similaridade de textos e o algoritmo KNN. Criou-se o índice de similaridade com os artigos selecionados no ponto 4.5. A decisão de um artigo ser considerado válido para a área "Continente", é feita com base no número de artigos já pertencentes à área Hipermercados Continente que surgem nos 5 mais similares, independentemente do seu grau de similaridade. O método de classificação automática de texto passa a ter a estrutura da Figura 4.13.

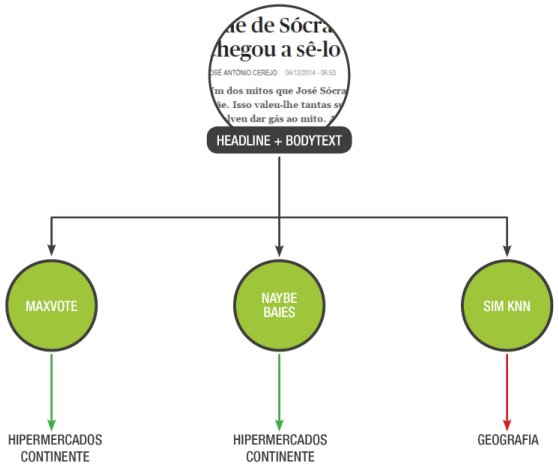


Figura 4.13- Simulação do método de classificação automática de um texto

Desta forma, para que um artigo seja considerado válido para a área, apenas terão de existir dois votos. Executando-se novamente os testes, obteve-se os valores apresentados no Gráfico 4.14.

Figura 4.14 – Resultados da área Continente nos seis períodos de tempo para fase 3 do sistema, com 3 classificadores

Com a implementação deste sistema de voto e analisando o Quadro 4.19 obteve-se um melhor compromisso entre as duas métricas, um *recall* de 87,64% e uma *precision* de 84,16%.

Quadro 4.19 – Evolução das métricas das várias fases do processo

	05.08 A 30.09		**05.08 A 16.12**	
	Recall	**Precision**	**Recall**	**Precision**
Similaridade através de frases	98,83%	66,40%	-	-
Similaridade através do texto completo	56,30%	93,66%	-	-
MaxVote	93,78%	85,76%	86,82%	81,94%
MaxVote + Naive Bayes	89,13%	86,59%	80,74%	84,40%
MaxVote + Naive Bayes + Sim KNN	93,70%	86,54%	87,64%	84,16%

5. Conclusão

Ao longo da investigação foram analisados alguns estudos, e todos eles partilham da mesma ideia chave: o volume de informação produzida para a Internet irá continuar a aumentar significativamente nos próximos anos. Neste sentido, a CISION Portugal sente a necessidade de implementar melhorias no seu processo de produção, de forma a ter capacidade para lidar com este fator, onde o tempo de entrega da informação é cada vez mais um elemento diferenciador perante os concorrentes, para além da qualidade do serviço prestado.

A CISION Portugal definiu como linha orientadora para o desenvolvimento do novo processo de indexação, a não utilização do histórico de artigos já validados pelas equipas de produção para as áreas em estudo, para avaliar a possibilidade de criar novas áreas sem haver a necessidade inicial de serem supervisionadas.

O sistema atual de produção está organizado por áreas, representando um tema. O processo de indexação dos conteúdos às áreas é feito por meio de palavras-chave combinadas com operadores booleanos.

Em áreas que abrangem temas bastante ambíguos, o número de artigos irrelevantes devolvidos para as equipas de produção é bastante alto, assim sendo a CISION Portugal com o objetivo de utilizar os mais recentes desenvolvimentos na área da NLP e de classificação automática de textos, decidiu verificar qual viabilidade da sua implementação nos processos de indexação nos sistemas de produção.

A primeira parte da investigação foi avaliar um método onde a definição das áreas por palavras-chave fosse substituída por meio da identificação e desambiguação de entidades existentes em artigos obtidos e armazenados pelo WISE. Este último é o sistema de monitorização de internet desenvolvido pela CISION Portugal.

Na investigação desenvolvida, as definições das entidades devem utilizar recursos de datasets públicos, como por exemplo, a Dbpedia.

As áreas iniciais selecionadas tinham um baixo grau de ambiguidade, ou seja, não estavam a ser supervisionadas pelas equipas de produção e passariam a ser definidas através da combinação booleana de entidades. Para efetuar a identificação e desambiguação de entidades no texto optou-se pelo sistema Dbpedia Spotlight devido ao facto de funcionar em Português e de não necessitar de qualquer tipo de treino.

Verificou-se também que o sistema Dbpedia Spotlight para a identificação e desambiguação de entidades, tinha uma ótima precisão nas áreas estudadas, mas falhava algumas ocorrências destas no texto, impossibilitando a indexação de alguns artigos às áreas.

De modo a encontrar uma solução para este problema, implementou-se um índice de similaridade por área. Este índice permite verificar a existência de semelhanças entre os artigos já indexados à área e os novos artigos, onde a anotação e desambiguação da entidade ou das entidades que façam parte da definição da área falhou. O índice foi preenchido com as frases dos artigos anotados corretamente pela DbPedia Spotlight, com as referências às entidades da área. Através deste método, os resultados foram melhorados, tendo-se obtido valores muito próximos aos do sistema atual. Com este método os artigos indexados para as áreas são bastante precisos, ou seja, fazem parte do âmbito da área. O ponto negativo é que existem alguns artigos que falham a indexação e pela análise dos resultados não é possível selecioná-los sem enviar um grande número de artigos irrelevantes para validação.

Com base nos resultados obtidos para as áreas com indexação automática, optou-se por validar o mesmo método para uma área que tivesse um elevado grau de ambiguidade: Hipermercados Continente. Após a aplicação do sistema descrito no parágrafo anterior, surgiram dois problemas: a Dbpedia Spotlight não fazia anotação e desambiguação correta do texto e o número de artigos identificados corretamente não foram suficientes para criar um bom índice similaridade. Para solucionar estas duas situações, optou-se por criar um corpus com duas classes, Hipermercados Continente e Geografia, para treinar um sistema de classificação automática de textos. Os artigos selecionados para a primeira classe tinham que ter entidades com um baixo grau de ambiguidade e um forte relacionamento com o Continente do grupo Sonae. Relativamente à classe Geografia foram selecionados artigos que estivessem de acordo com a definição gramatical da palavra, mas sempre através de entidades, e que não incluíssem nenhuma entidade que definia os Hipermercados Continente. Após a realização de vários testes e de tipos de implementação verificou-se que os melhores resultados foram obtidos através do método de votação.

O método de votação é composto por três classificadores de diferentes tipos:

- **MaxVote**: é um classificador que combina as respostas dos classificadores Naive Bayes, Decision Tree e o MaxEnt. O resultado obtido é a classe que possuir maior número de votos internamente;
- **Naive Bayes**: o classificador foi treinado através de um método diferente do mesmo tipo de classificador utilizado pelo MaxVote;
- **KNN + Similaridade entre documentos**: compara um novo artigo com os artigos das duas classes referidas anteriormente e verifica qual a classe vencedora nos cinco documentos similares.

A resposta devolvida por parte do sistema descrito será a classe que obter o maior número de votos. Com esta implementação conseguiu-se obter um *recall* de 87,64% e uma *precision* de 84,16%.

A métrica *precision* obtida através do sistema proposto é um pouco baixa, contudo não se pode esquecer que todo o sistema foi treinado sem qualquer tipo de supervisionamento e sem usar o histórico validado pelas equipas de produção para a área dos Hipermercados Continente. O caso da utilização do histórico na implementação deste protótipo em produção, permite-nos acreditar que estas duas métricas irão certamente subir para valores acima dos 90%, dado que se identificaram vários artigos selecionados para a classe Hipermercados Continente, onde o seu relacionamento é bastante baixo.

A investigação demonstrou que a abordagem proposta neste documento, quando aplicada em áreas de indexação automática e para o caso terem um baixo grau de ambiguidade, funciona e têm-se valores quase idênticos aos que a CISION Portugal tem atualmente no sistema de produção. A grande vantagem do método proposto neste documento é que as indexações têm um elevado grau de certeza, com valores muito próximo dos 100%, ou seja, as indexações feitas para áreas são corretas.

Relativamente às áreas com um elevado grau de ambiguidade, a investigação não foi muito profunda, por isso não se pode afirmar taxativamente que o sistema de votação irá funcionar corretamente, sem que haja um supervisionamento na seleção dos artigos para o corpus. Contudo, mostra um caminho possível para ser aplicado em áreas novas criadas num sistema deste tipo.

Outro dado relevante identificado durante a investigação para a área dos Hipermercados Continente é que atualmente a equipa de validação por cada sete dias, valida em média 692 artigos. Na eventualidade de se optar pela utilização da pesquisa c*ase sensitive* (sensível a maiúsculas), consegue-se reduzir aproximadamente para um quarto os resultados propostos para validação. Outra proposta é a utilização do método de classificação por votação, que permite reduzir para um quinto das propostas a validar. Uma outra alternativa para a redução de propostas, é considerar irrelevantes todos os artigos onde o classificador foi unânime em classificar como geografia, mas teríamos uma taxa de erro de 6%.

Tendo em consideração que no dataset, o objeto de estudo tinha um total de 391 210 documentos de seis períodos de tempo distintos e os resultados obtidos, verifica-se que o caminho a percorrer pela CISION Portugal deverá passar pela implementação de um sistema idêntico ao apresentado nesta investigação, mas analisando área a área.

Para trabalho futuro será importante numa primeira fase fazer um levantamento área a área para verificar a possibilidade das pesquisas serem *case sensitive*. O segundo passo será identificar as áreas que tenham baixa ambiguidade e aferir quais poderão ser transferidas para esta abordagem. Para as áreas que necessitam de um sistema de classificação por votação deve-se utilizar, sempre que possível, o histórico existente na base de dados da CISION Portugal para criar o corpus de treino.

Durante o período desta investigação foi desenvolvido um interface gráfico protótipo de gestão do sistema proposto na investigação, no âmbito da realização de um estágio curricular da Licenciatura de Engenharia de Informática e Sistemas do Instituto Superior de Engenharia de Coimbra.

O interface desenvolvido permite fazer a gestão de acessos por meio de identificação dos utilizadores e da gestão das áreas através da utilização dos recursos da Wikipedia, visualizando os resultados por área e os detalhes dos artigos.

Para o caso do sistema proposto no presente documento ser implementado nos sistemas de produção da CISION Portugal, o protótipo desenvolvido deverá ser analisado e melhorado para permitir a sua utilização por parte dos colaboradores da empresa.

6. Referências Bibliográficas

algorithm, k-nearest neighbors. 2014. k-nearest neighbors algorithm. *Wikipedia.* [Online] 24 de 11 de 2014. http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

Alves, Alexandra Isabel Magalhães. 2010. *Modelo de representação de texto mais adequado à classificação* . 2010.

Bekkerman, Ron e Allan, James. 2004. *Using Bigrams in Text Categorization.* s.l. : CIIR Technical Report, 2004.

Berry, Michael W. e Kogan, Jacob. 2010. *Text Mining - Applications and Theory.* 2010.

Caropreso, Maria Fernanda, Matwin, Stan e Sebastiani, Fabrizio. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. [autor do livro] Amita G. Chin. *Text Databases and Document Management: Theory and Practice.* s.l. : Idea Group Publishing, 2001.

David Reby, Sovan Lek, Ioannis Dimopoulos, Jean Joachim, Jacques Lauga, Stéphane. 1997. *Artificial neural network as a classification method in the behavioural sciences.* 1997.

Duarte, Júlio César. 2009. *O Algoritmo Boosting at Start e suas Aplicações.* 2009.

Escudeiro, Nuno Filipe Fonseca Vasconcelos. 2004. *"Automatic Web Resource Compilation Using Data Mining" - MSc Thesis on Data Analysis and Decision Support Systems.* 2004.

Fayyad, Usama, Piatetsky-Shapiro, Gregory e Smyth, Padhraic. 1996. *From Data Mining to Knowledge Discovery in Databases.* 1996. pp. 37-54.

Feature subset selection in text-learning. Mladenic, Dunja. 1998. s.l. : ECML98, 1998. 10th European Conference on Machine Learning.

Feldman, Ronen e Sanger, James. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* s.l. : Cambridge University Press, 2007.

Fonseca, Dário. 2013. Let's be social. 2013.

Freitas, Ana Teresa. 2002. Introdução à Biologia Molecular. *Inesc.* [Online] 2002.

Gama, João. 2002. *Árvores de Decisão.* 2002.

Gomes, George Alex Fernandes. 2012. *Eu-Tu: o emprego da classificação automática de mensagens em fóruns eletrônicos de discussões para análise do processo de ensino e aprendizagem centrado em interações.* 2012.

Gomes, Helder Joaquim Carvalheira. 2012. *Text Mining: Análise de Sentimentos na classificação de notícias.* 2012.

Gonçalves, Teresa e Quaresma, Paulo. 2005. *Evaluating preprocessing techniques in Text Classification problem.* s.l. : Departamento de Informática, Universidade de Évora, 2005.

Hsinchun Chen, Feiyue Wang,Christopher C. Yang. 2006. *Intelligence and Security Informatics: International Workshop.* 2006.

Jackson, P. e Moulinier, I. 2002. *Applications: Text Retrieval, Extraction, and Categorization Applications: Text Retrieval, Extraction, and Categorization.* 2002.

Lobo, V. 2010. *Sistemas de Apoio à Decisão– Árvores de decisão.* 2010.

Lobo, Victor. 2010. *Arvores de decisão.* 2010.

MAHESH , T R, SURESH, M B e VINAYABABU, M . 2009. *TEXT MINING: ADVANCEMENTS, CHALLENGES AND FUTURE DIRECTIONS.* 2009. pp. 61-65.

Manning, Christopher D., Raghavan, Prabhakar e Schütz, Hinrich. 2009. Naive Bayes text classification. *Introduction to Information Retrieval.* [Online] Cambridge University Press, 2009. http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.

Marktest. 2014. Mais portugueses online. *Marktest - Noticias.* [Online] Outubro de 2014. http://www.marktest.com/wap/a/n/id~1e07.aspx.

Mendes, Pablo N., et al. 2011. *DBpedia Spotlight: Shedding Light on the Web of Documents.* 2011.

MITCHELL, T.M. 1997. *Machine Learning.* s.l. : McGraw Hill, 1997.

Nielsen. 2012. *STATE OF THE MEDIA: THE SOCIAL MEDIA REPORT 2012.* 2012.

Nigamy, Kamal, Lafferty, John e McCallumzy, Andrew. 1999. *Using Maximum Entropy for Text Classi.* 1999.

Pereira, João José Rodrigues. 2005. *Modelos de data mining para multi-previsão: aplicação à medicina intensiva.* 2005.

Perkins, Jacob. 2010. *Python Text Processing with NLTK 2.0 Cookbook.* Mumbai : Packt Publishing, 2010.

RATNAPARKHI, A. 1997. *A Simple Introduction to Maximum Entropy Models for Natural.* 1997.

Reis, Laudo e Romero, Roseli Aparecida Francelin. 2008. *Inteligência Computacional Aplicada à Análise de Risco no Contexto do Tratado da Basiléia.* 2008.

Rodrigues, Jorge Nascimento e Devezas, Tessaleno. 2007. *Portugal – O Pioneiro da Globalização.* s.l. : Centro Atlântico, 2007.

Rolim, Pedro Gonçalo Jorge. 2011. *NovaIntell – Projecto de Text Mining para a língua portuguesa numa empresa de Gestão de Informação e Conhecimento.* 2011.

Santos, António Paulo Gomes dos. 2008. *Classificação multi-etiqueta hierárquica de textos segundo a taxonomia ACM.* 2008.

Sebastiani, Fabrizio. 2002. *Machine learning in automated text categorization.* s.l. : ACM Computing, 2002.

Spark-Jones e K., e P. Willett. 1997. *Readings in Information Retrieval. : Morgan.* San Francisco : Morgan Kaufmann, 1997.

Statsoft. Naive Bayes Classifier. *Statsoft.* [Online] [Citação: 09 de 12 de 2014.] http://www.statsoft.com/textbook/naive-bayes-classifier.

Sullivan, D. 2001. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales.* New York, NY, USA : John Wiley & Sons, Inc, 2001.

Teixeira, Paulo Jorge Pereira. 2014. *BrainSleep Sistema móvel para deteção de estados de sonolência baseado em ondas cerebrais* . 2014.

Thuraisingham, Bhavani. 1998. *Data Mining: Technologies, Techniques, Tools, and Trends.* s.l. : CRC Press, 1998.

Trigo, Luís Manuel Pimentel. 2010. *Análise de Proximidade entre Investigadores de Alguns Centros de I&D da Universidade do Porto usando Text Mining sobre Bases de Dados Bibliográficas.* 2010.

Trigueiros, Duarte. 1991. *As Árvores de Decisão.* 1991.

Um estudo e apreciação sobre algoritmos de stemming para a Língua Portuguesa. Chaves, Marcirio Silveira. 2003. Cartagena de Indias, Colombia : s.n., 2003. IX Jornadas Iberoamericanas de Informática.

Universe, EMC Digital. 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. *Emc.com.* [Online] 2014.

Vizzuality, Hyperakt e. 2011. Evolution of the Web. *Evolution of the Web.* [Online] Hyperakt e Vizzuality, 2011. http://www.evolutionoftheweb.com/.

Wikipedia. Latent semantic indexing. *Wikipedia.* [Online] [Citação: 9 de 12 de 2014.] http://en.wikipedia.org/wiki/Latent_semantic_indexing.

X. Wu, V. Kumar, J Ross Quinlan, J. Ghosh, Q. Yang, H.Motoda. 2007. *Top 10 algorithms in data mining.* 2007.

Yang, Yiming. 1999. *An Evaluation of Statistical Approaches to Text Categorization.* s.l. : Journal of Information Retrieval, 1999.

Yhat. 2013. Intuitive Classification using KNN and Python. *Yhat Blog* . [Online] 25 de 7 de 2013. http://blog.yhathq.com/posts/classification-using-knn-and-python.html.

7. Anexos

7.1. Anexo 1 - Protótipo

Para demonstração das conclusões e das tecnologias utilizadas na investigação desenvolveu-se um protótipo que foi dividido em duas secções: similaridade de artigos e segmentação de artigos através do reconhecimento de entidades.

Após as conclusões obtidas neste estudo devido à utilização de um índice de similaridade para a resolução das falhas de identificação e desambiguação de entidades por parte da DbPedia Spotlight, a CISION Portugal verificou a potencialidade deste método e decidiu implementar esta técnica estudada nos seus sistemas de produção, com o objetivo de diminuir no imediato o volume de notícias validadas pela equipa de conteúdos da internet.

A CISION Portugal definiu que o índice incluiria todos os artigos de internet de idioma português captados nos últimos dois dias e seria atualizado à medida que os artigos fossem captados e inseridos no sistema de produção. Com a introdução deste método, a empresa consegui diminuir diariamente em média 16% o volume de artigos enviados para a equipa de validação de conteúdos de internet. Contudo durante a fase de implementação e de testes verificou-se que este número poderá aumentar para valores compreendidos entre os 20% e os 25%, mas para atingir este volume será necessário realizar alguns ajustes nos processos de indexação a jusante deste passo. A demonstração de uma parte do processo atrás descrito, encontra-se na secção “Similaridade” do protótipo.

O protótipo é composto por um portal web desenvolvido através ASP.NET MVC, uma base de dados em MySQL, um índice SOLR e por um web service desenvolvido em Python. O esquema geral do protótipo encontra-se esquematizado na Figura 7.2.

O portal web, ver Figura 7.1, permite que um utilizador possa optar por verificar o grau de similaridade entre artigos ou por obter a segmentação de artigos através da identificação de entidades. Para o armazenamento dos artigos usados na demonstração do protótipo foi utilizada uma base de dados MySQL. Os métodos de similaridade e segmentação de artigos estão disponíveis através do web service desenvolvido em Python, que por sua vez comunica com o web service disponibilizado pela DbPedia Spotlight, para a obtenção das entidades. Os artigos, e as entidades extraídas foram armazenados num índice SOLR.

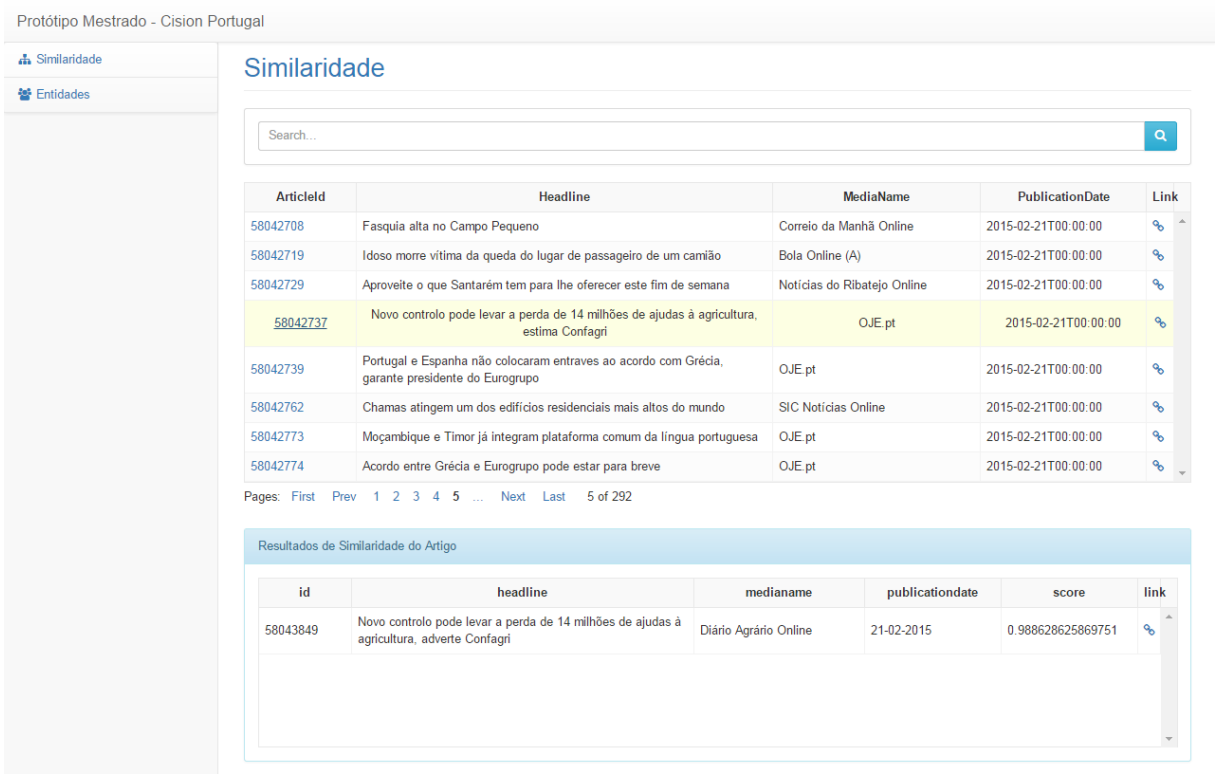


Figura 7.1 – Portal web do protótipo

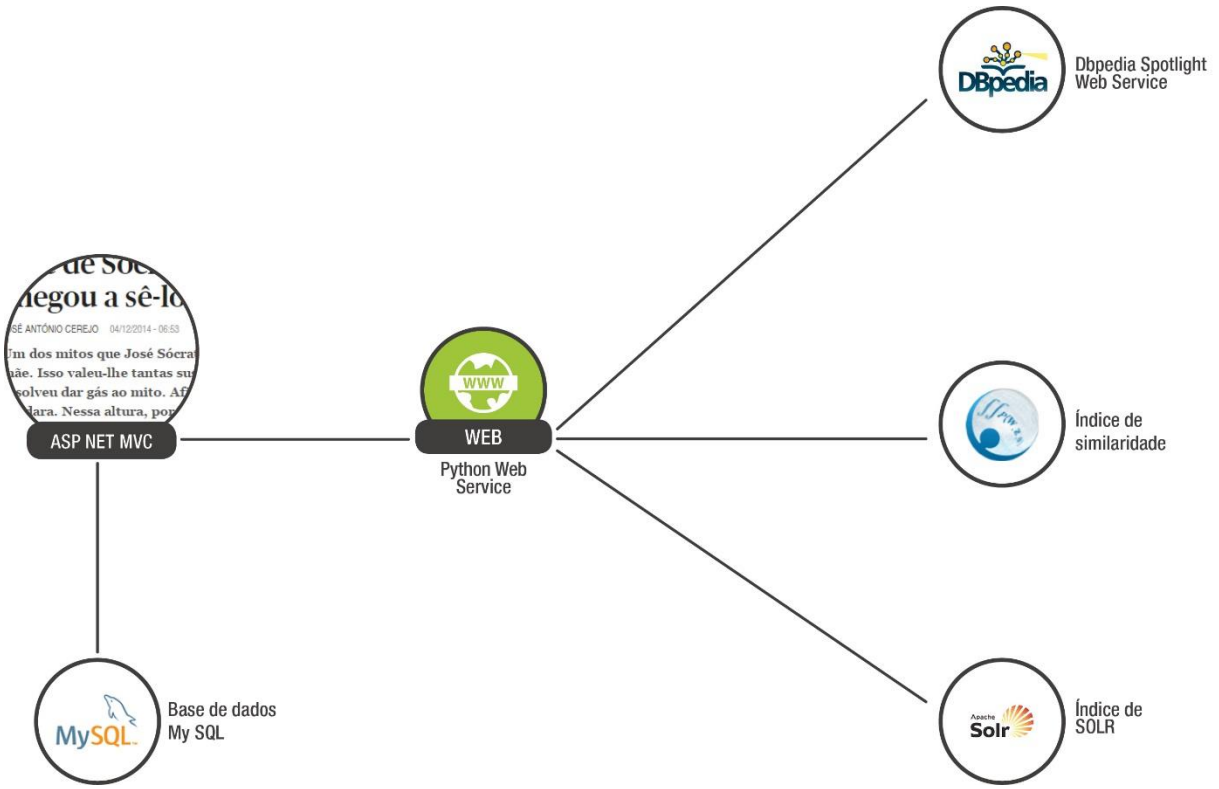


Figura 7.2 – Esquema geral do protótipo

7.1.1. Casos de Uso e Diagramas de Atividades

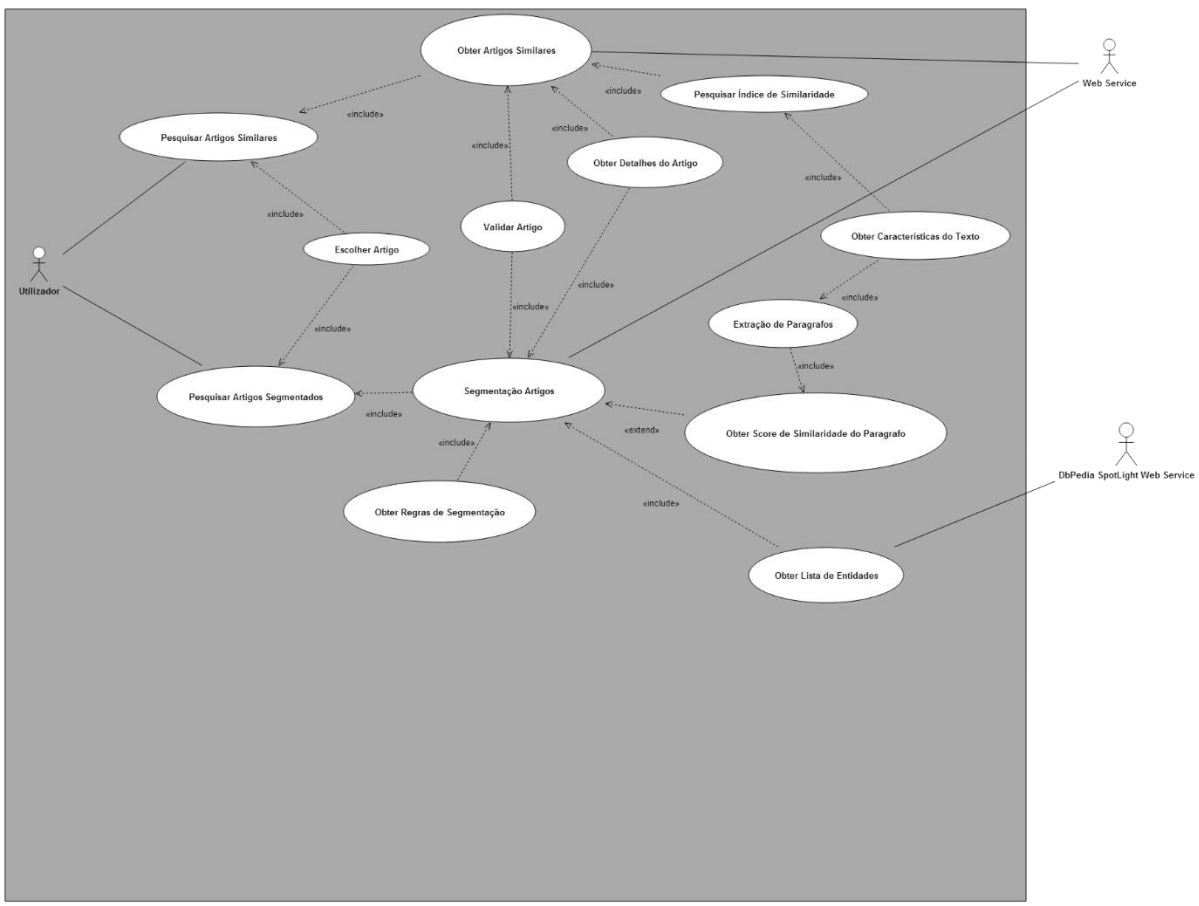


Figura 7.3 – Casos de Uso

Caso de Uso: Pesquisar Artigos Similares

Fluxo de Eventos:

1. O utilizador liga-se à plataforma;
2. O utilizador escolhe a opção “Artigos Similares”
3. O sistema apresenta a lista de artigos;
4. O utilizador seleciona um artigo;
5. Faz caso de uso “Obter Artigos Similares”
6. O sistema apresenta os artigos similares;
7. O caso de uso termina com sucesso.

Cenários Alternativos:

5.a. O caso de uso “Obter Artigos Similares” termina sem sucesso

5.a.1. O sistema devolve uma lista de artigos vazia

5.a.2. O caso de uso termina sem sucesso.

6.a. A lista de artigos similares devolvidos é vazia

6.a.1. O sistema informa o utilizador através de uma mensagem.

6.a.2. O caso de uso termina com sucesso.

Diagrama de Atividade:

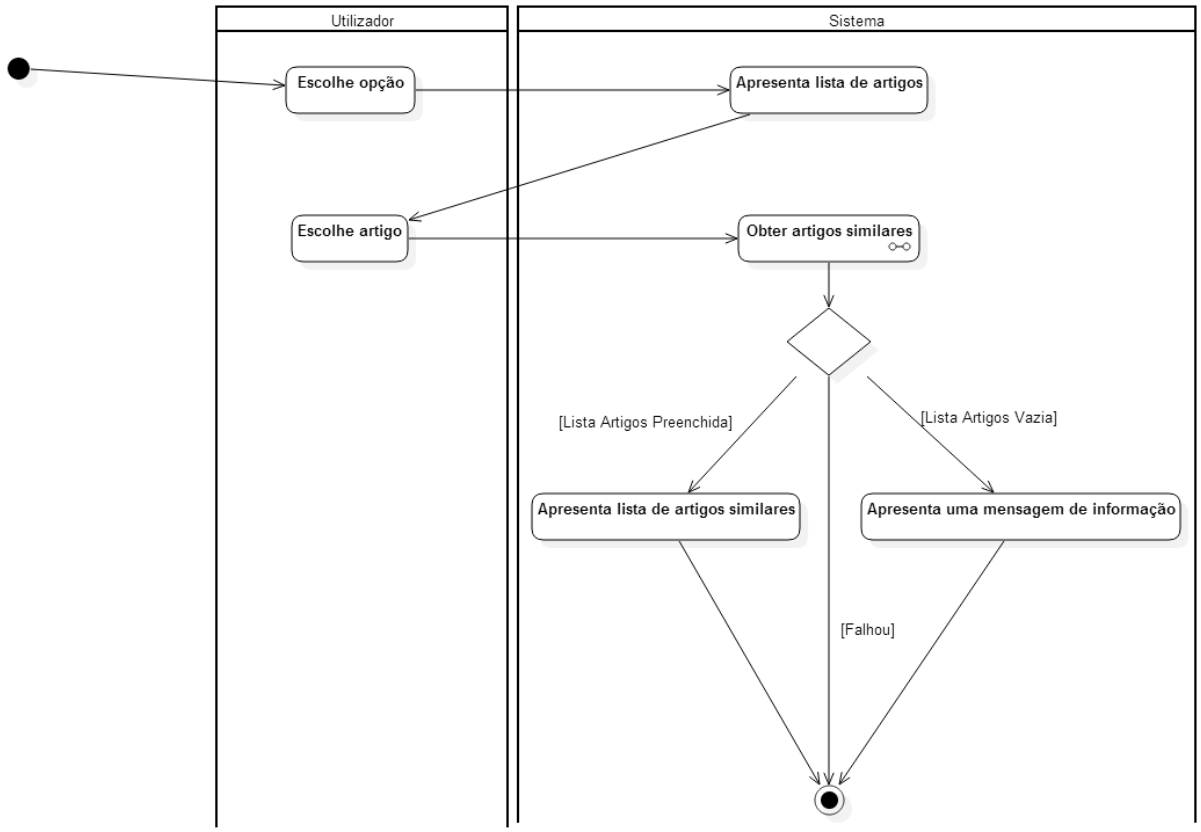


Figura 7.4 - Diagrama de Atividade "Pesquisar Artigos Similares"

Caso de Uso: Obter Artigos Similares

Fluxo de Eventos:

1. O sistema recebe um id de um artigo
2. Faz o caso de uso “Obter Detalhes do Artigo”
3. Faz o caso de uso “Validar Artigo”
4. Faz o caso de uso “Pesquisar Índice de Similaridade”
5. O sistema devolve a lista de artigos similares
6. O caso de uso termina com sucesso

Cenários Alternativos:

2.a. O caso de uso “Obter Detalhes do Artigo” não encontra o artigo

2.a.1. O sistema devolve um artigo vazio

2.a.2. Faz o caso de uso “Validar Artigo”

2.b. O caso de uso “Obter Detalhes do Artigo” termina sem sucesso

2.b.1. O sistema devolve um artigo vazio

2.b.2. Faz o caso de uso “Validar Artigo”

3.a. O artigo é inválido

3.a.1. O sistema devolve uma lista de artigos similares vazia

3.a.2. O caso de uso termina com sucesso

3.b. O caso de uso “Validar Artigo” termina sem sucesso

3.b.1. O sistema devolve uma lista de artigos similares vazia

3.b.2. O caso de uso termina com sucesso

4.a. O caso de uso “Pesquisar Índice de Similaridade ” termina sem sucesso

4.a.1. O sistema devolve uma lista de artigos similares vazia

4.a.2. O caso de uso termina com sucesso.

Diagrama de Atividade:

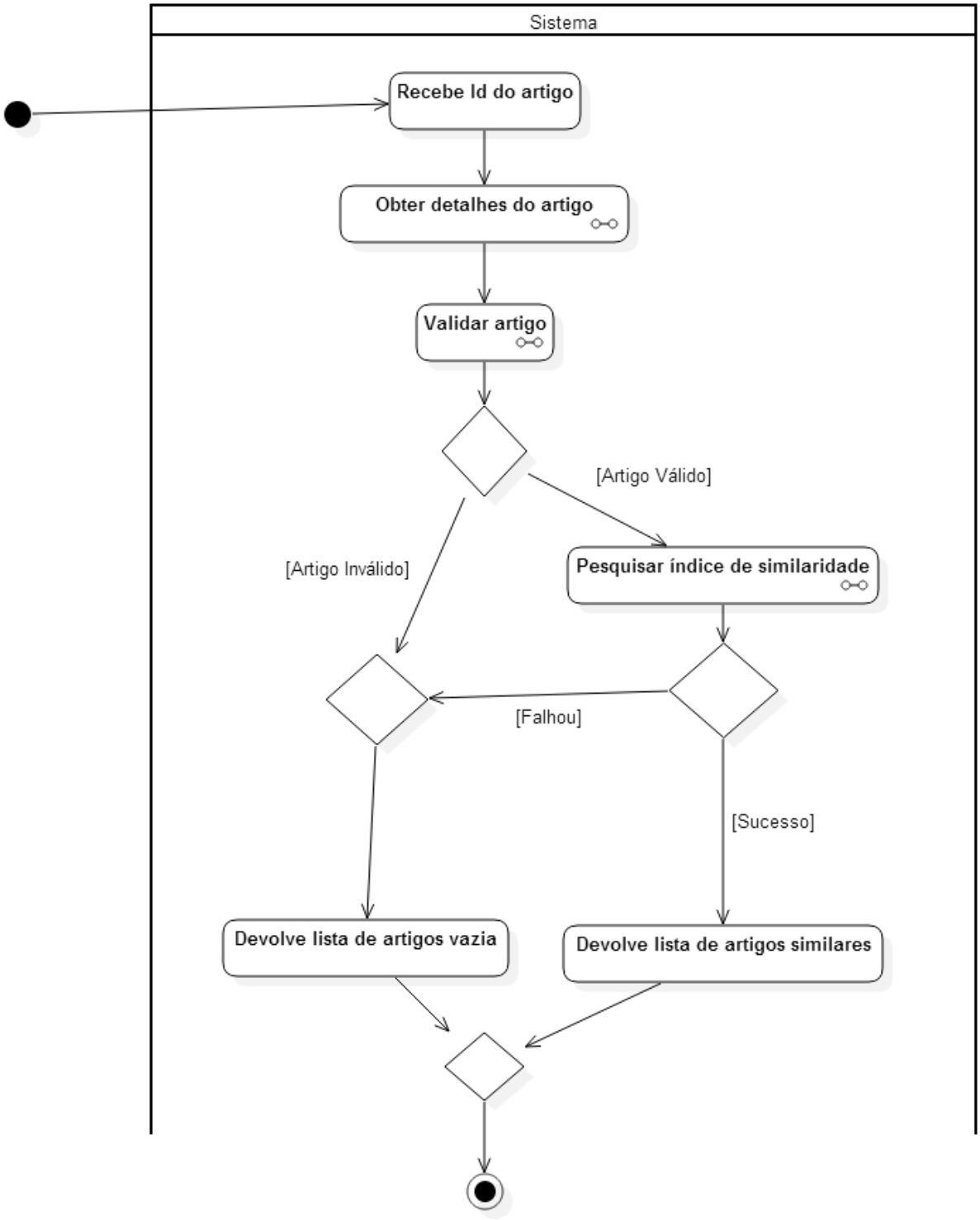


Figura 7.5 – Diagrama de Atividade "Obter Artigos Similares"

Caso de Uso: Validar Artigo

Fluxo de Eventos:

1. Recebe Texto;
2. O sistema valida se o texto contém mais que 50 caracteres;
3. O sistema valida se o texto está escrito no idioma Português;
4. O caso de uso termina com sucesso;

Cenários Alternativos:

2.a. O tamanho do texto do artigo é inferior a 50 caracteres:

2.a.1. O sistema retorna artigo inválido

2.a.2. O caso de uso termina com sucesso

2.b. O tamanho do texto do artigo é maior ou igual a 50 caracteres:

2.b.1. O sistema continua para o passo 3

3.a. O idioma do texto do artigo é diferente de Português:

3.a.1. O sistema retorna artigo inválido

3.a.2. O caso de uso termina com sucesso

3.b. O idioma do texto do artigo é Português:

3.b.1. O sistema retorna artigo como válido

3.b.2. O caso de uso termina com sucesso

Diagrama de Atividade:

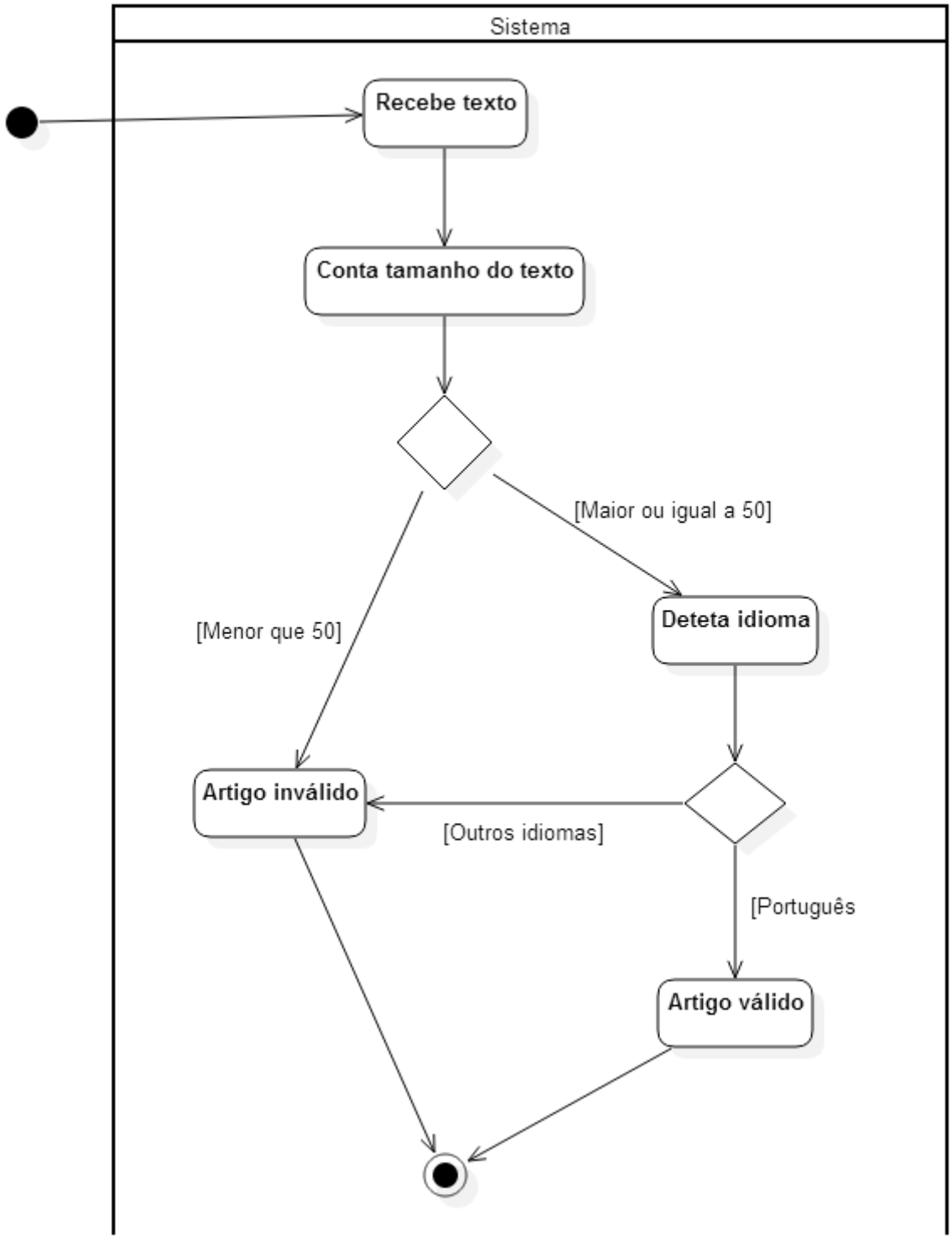


Figura 7.6 – Diagrama de Atividade "Validar Artigo"

Caso de Uso: Pesquisar Índice de Similaridade

Fluxo de Eventos:

1. Recebe Texto;
2. Faz o caso de uso “Remover StopsWords”;
3. O sistema pesquisa quais os artigos mais similares com o texto em análise;
4. O sistema indexa o artigo ao índice de similaridade;
5. O sistema devolve a lista de artigos similares;
6. O caso de uso termina com sucesso;

Cenários Alternativos:

2.a. O caso de uso “Remover StopsWords” termina sem sucesso:

2.a.1. O sistema devolve uma lista de palavras vazia;

2.a.2. O sistema devolve uma lista de artigos similares vazia;

2.a.3. O caso de uso termina com sucesso;

3.a. O sistema não encontra artigos similares:

3.a.1. O sistema retorna uma lista de artigos vazia

3.a.2. O caso de uso termina com sucesso

3.b. O sistema falha na pesquisa de artigos similares:

3.b.1. O sistema retorna uma lista de artigos vazia

3.b.2. O caso de uso termina com sucesso

Diagrama de Atividade:

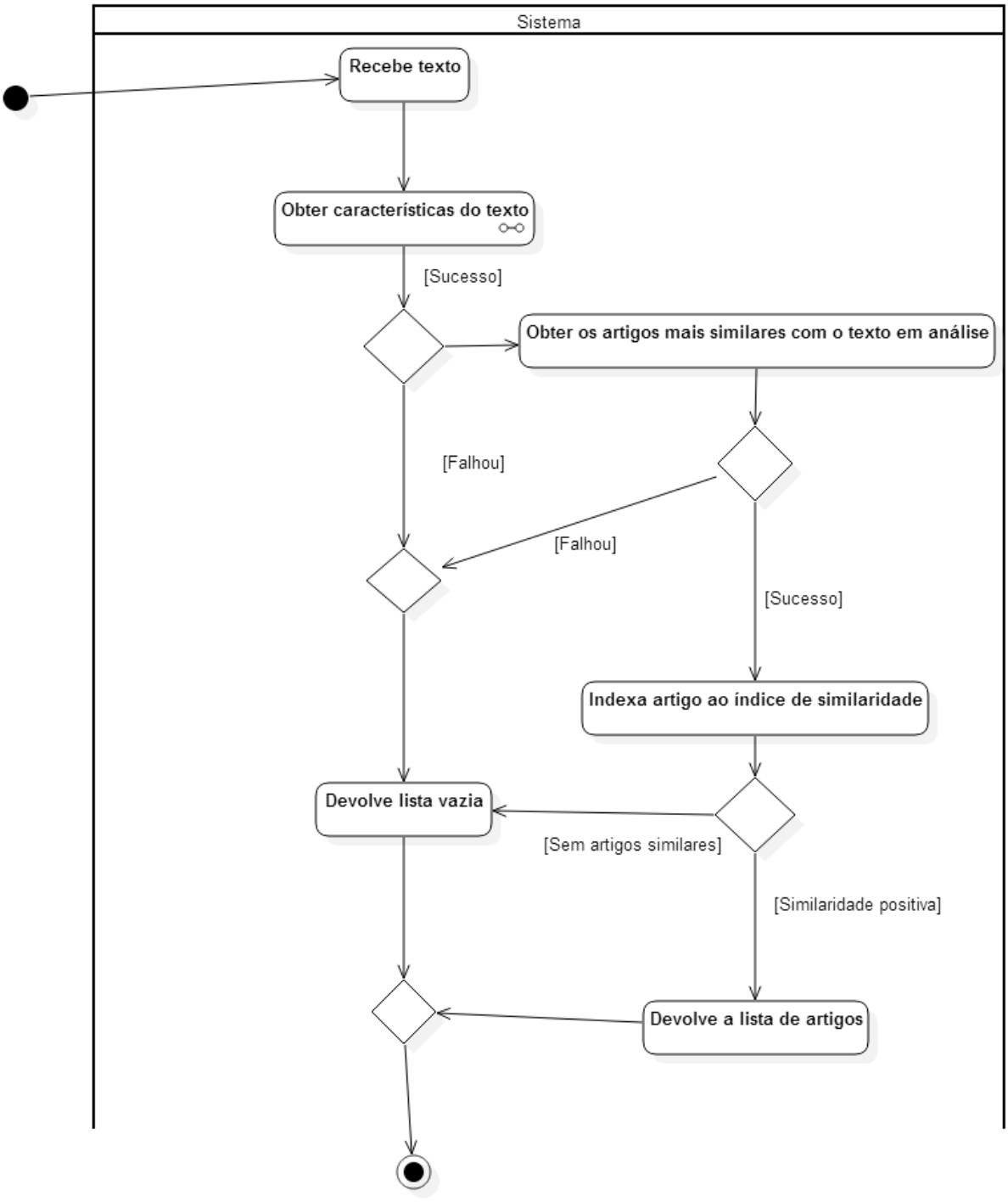


Figura 7.7 – Diagrama de Atividade "Pesquisar Índice Similaridade"

Caso de Uso: Obter Detalhes do Artigo

Fluxo de Eventos:

1. O sistema recebe um id de um artigo
2. O sistema pesquisa o id no índice SOLR
3. O sistema devolve detalhes do artigo
4. O caso de uso termina com sucesso

Cenários Alternativos:

2.a. O sistema não encontra o artigo

2.a.1. O sistema devolve os detalhes do artigo vazios;

2.a.2. O caso de uso termina com sucesso;

2.b. Falha a pesquisa no índice SOLR

2.a.1. O sistema devolve os detalhes do artigo vazios;

2.a.2. O caso de uso termina com sucesso;

Diagrama de Atividade:

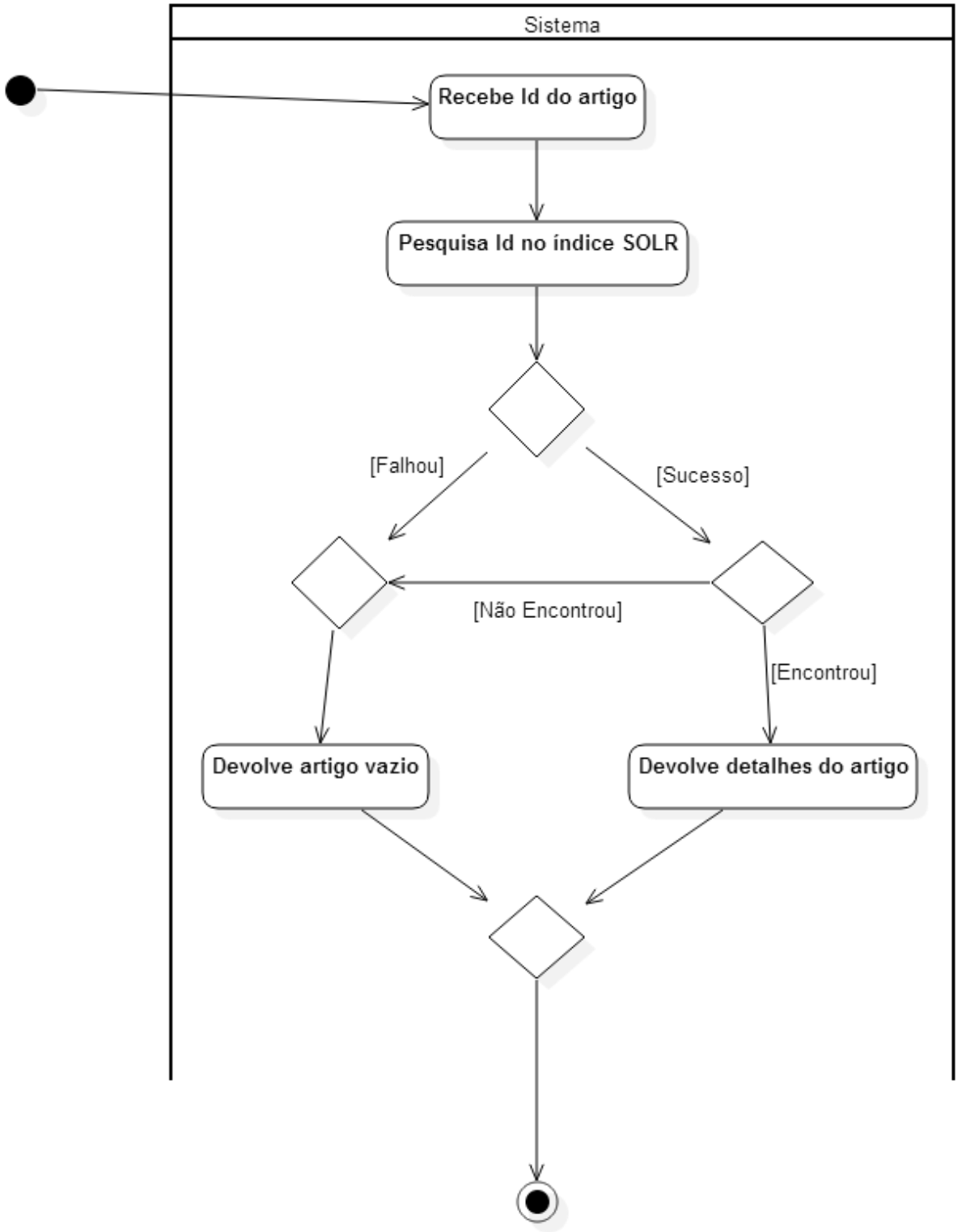


Figura 7.8 – Diagrama de Atividade "Obter Detalhes do Artigo"

Caso de Uso: Pesquisar Artigos Segmentados

Fluxo de Eventos:

1. O utilizador liga-se à plataforma;
2. O utilizador escolhe a opção “Artigos Segmentados”
3. O sistema apresenta a lista de artigos;
4. O utilizador seleciona um artigo;
5. Faz caso de uso “Segmentação artigos por área”
6. O sistema apresenta as áreas associadas ao artigo;
7. O caso de uso termina com sucesso.

Cenários Alternativos:

5.a. O caso de uso “Segmentação Artigos por Área” devolve artigo sem interesse

5.a.1. O sistema apresenta o artigo sem segmentação.

5.a.2. O caso de uso termina com sucesso.

5.b. O caso de uso “Segmentação Artigos por Área” devolve artigo com interesse

5.b.1. O sistema apresenta o artigo com segmentação.

5.b.2. O caso de uso termina com sucesso.

Diagrama de Atividade:

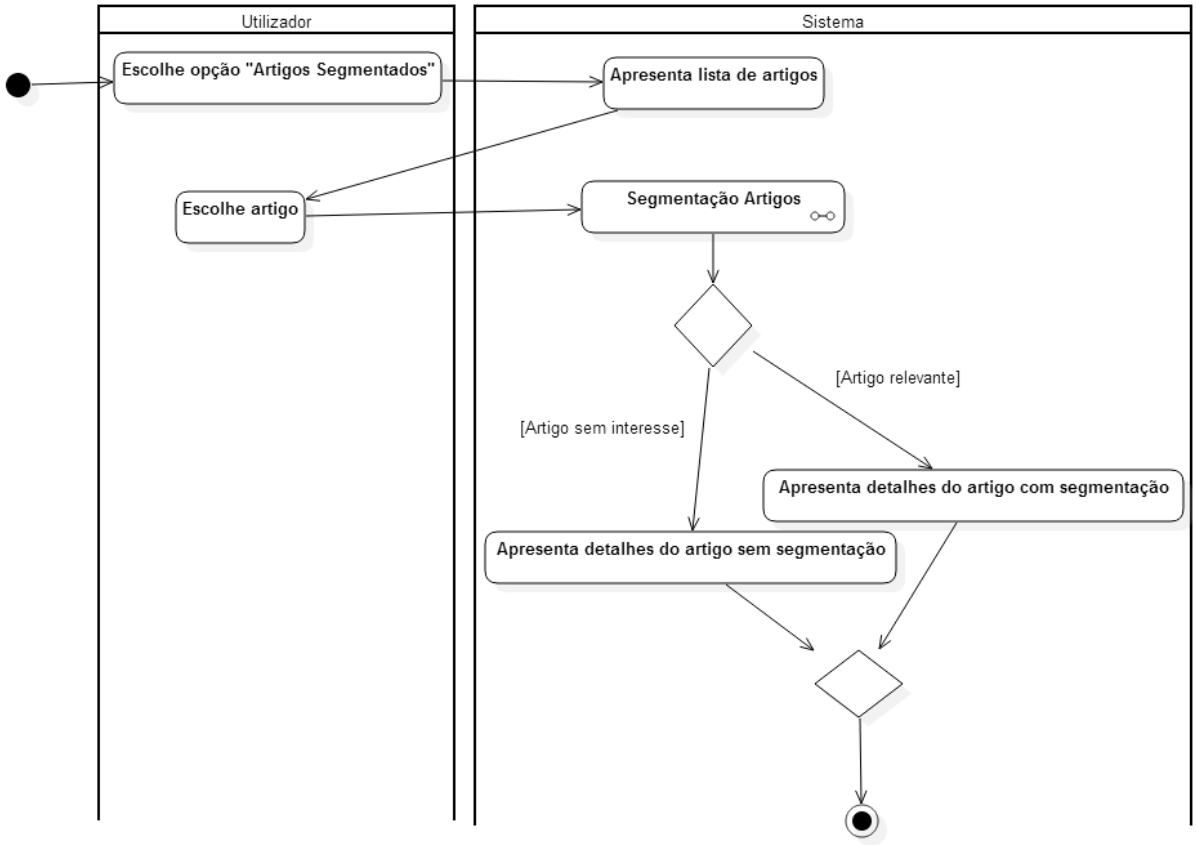


Figura 7.9 – Diagrama de Atividade "Pesquisar Artigos Segmentados"

Caso de Uso: Obter Regras de Segmentação

Fluxo de Eventos:

1. O sistema lê as regras de segmentação de cada área da base de dados
2. O sistema devolve lista de regras de segmentação
3. O caso de uso termina com sucesso

Cenários Alternativos:

2.a. O sistema não encontra regras de segmentação

2.a.1. O sistema devolve uma lista de regras vazia

2.a.2. O caso de uso termina sem sucesso;

2.b. Falha a leitura das regras de segmentação

2.b.1. O sistema devolve uma lista de regras vazia

2.b.2. O caso de uso termina sem sucesso

Diagrama de Atividade:

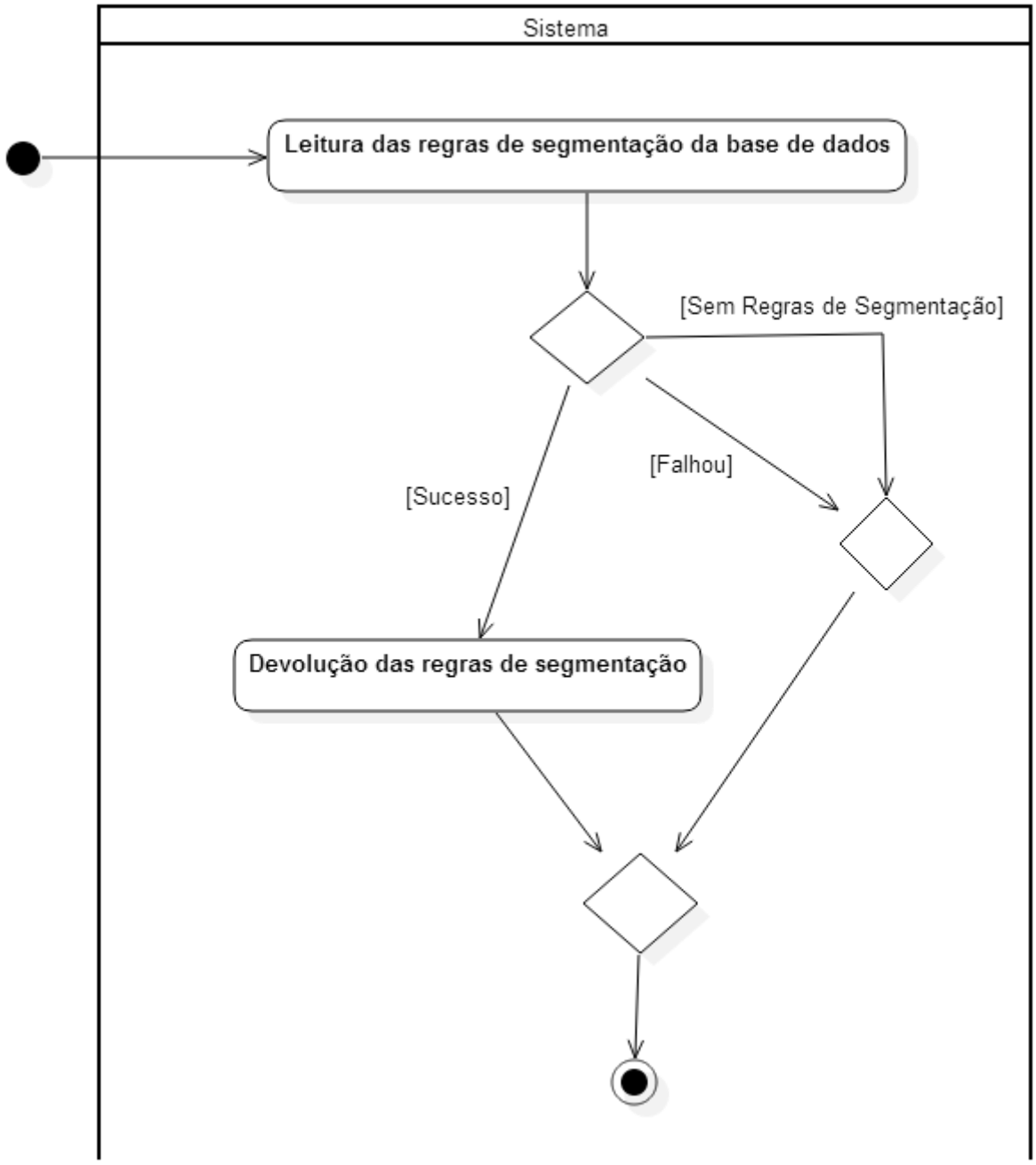


Figura 7.10 - Diagrama d Atividade "Obter Regras de Segmentação"

Caso de Uso: Obter Lista de Entidades

Fluxo de Eventos:

1. O sistema recebe texto do artigo
2. Envio do texto para o Web Service da Dbpedia Spotlight
3. O sistema retorna a lista de entidades devolvida pelo web service;
4. O caso de uso termina com sucesso

Cenários Alternativos:

2.a. O sistema falhou a chamada ao web service

2.a.1. O sistema retorna a lista de entidades vazia;

2.a.2. O caso de uso termina com sucesso;

Diagrama de Atividade:

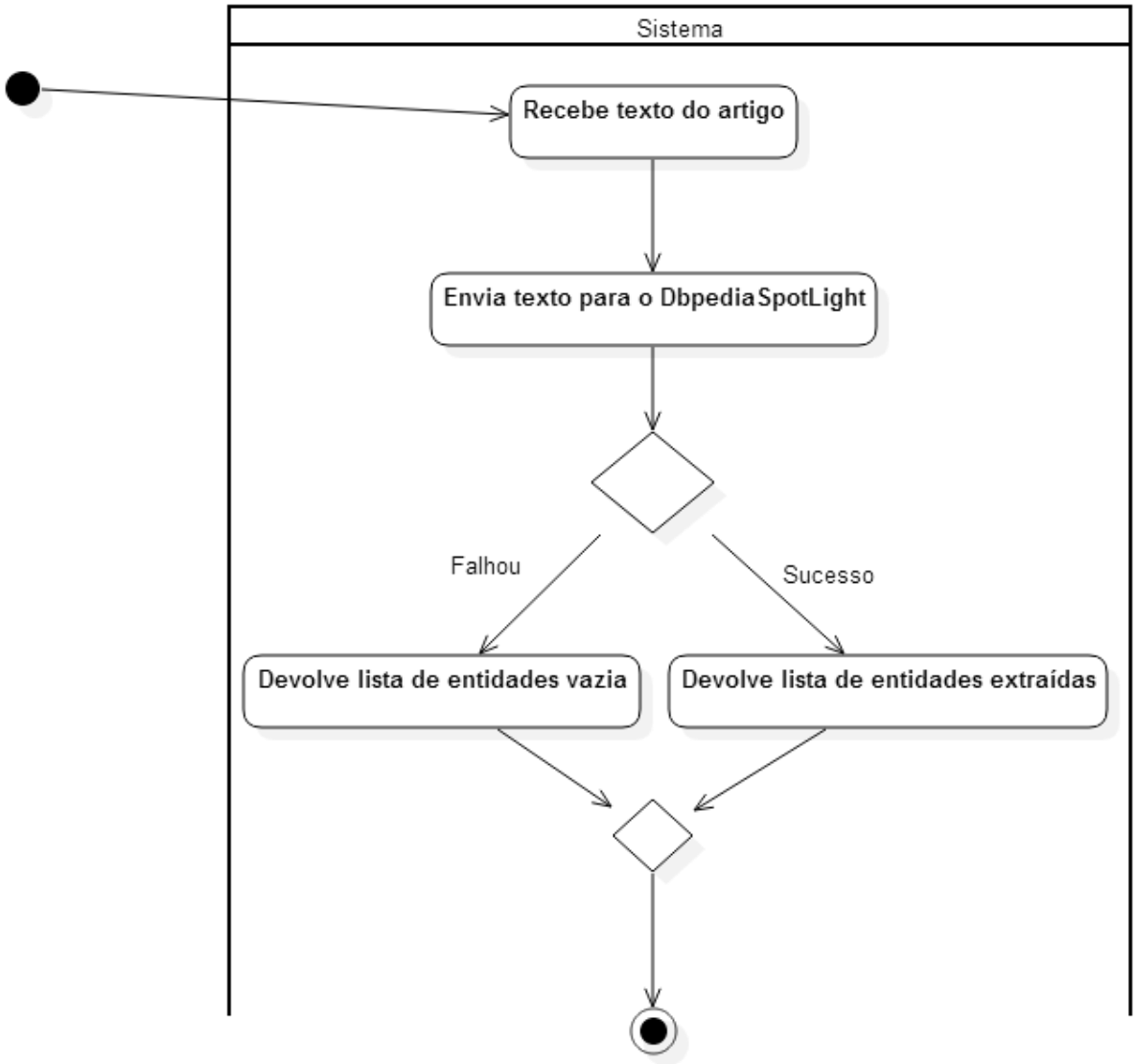


Figura 7.11 – Diagrama de Atividade "Obter Lista de Entidades"

Caso de Uso: Obter Características do Texto

Fluxo de Eventos:

1. O sistema recebe texto
2. Remoção das tags de HTML
3. Decomposição do texto em palavras
4. Remoção das stopwords
5. Stemming
6. O sistema devolve uma lista de palavras
7. O caso de uso termina com sucesso

Diagrama de Atividade:

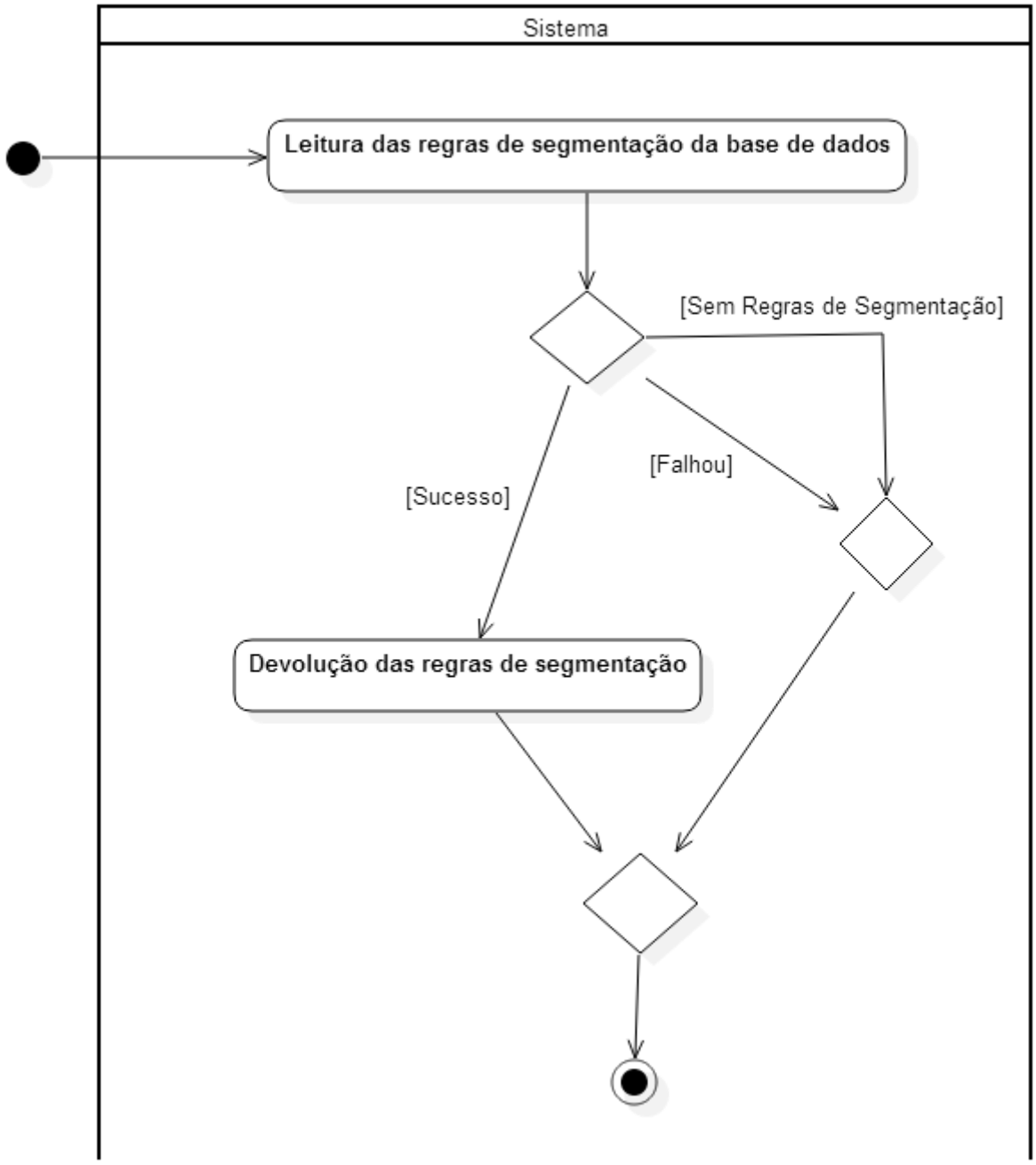


Figura 7.12 – Diagrama de Atividade "Obter Características do Texto"

Caso de Uso: Extração de Parágrafos

Fluxo de Eventos:

1. O sistema recebe texto
2. Separação do texto em parágrafos
3. Extração dos parágrafos com referências às entidades
4. Faz caso de uso “Obter características do texto”
5. O sistema devolve lista de parágrafos
6. O caso de uso termina com sucesso

Cenários Alternativos:

3.a. Não existem referências

3.a.1. O sistema retorna a lista de parágrafos vazia

3.a.2. O caso de uso termina com sucesso

4.a. O caso de uso “Obter características do texto” falha

4.a.1. O sistema retorna a lista vazia

4.a.2. O caso de uso termina com sucesso

Diagrama de Atividade:

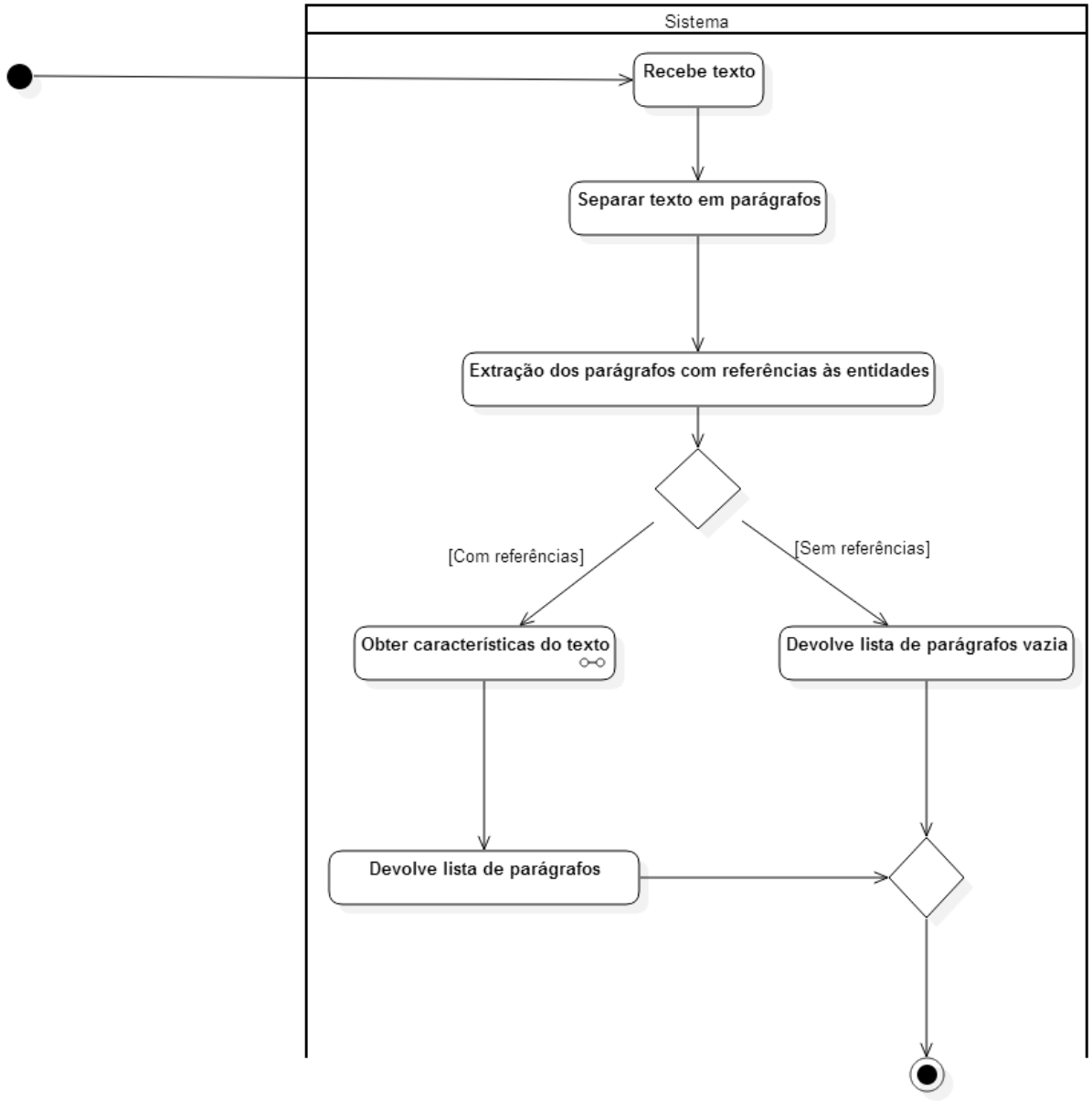


Figura 7.13 – Diagrama de Atividade "Extração de Parágrafos"

Caso de Uso: Obter Score de Similaridade do Paragrafo

Fluxo de Eventos:

1. O sistema recebe parágrafo
2. Pesquisar índice pelos parágrafos mais similares
3. Calcular “full score” de similaridade
4. Calcular “10% score” de similaridade
5. O sistema devolve “full score” e “10% score”
6. O caso de uso termina com sucesso

Diagrama de Atividade:

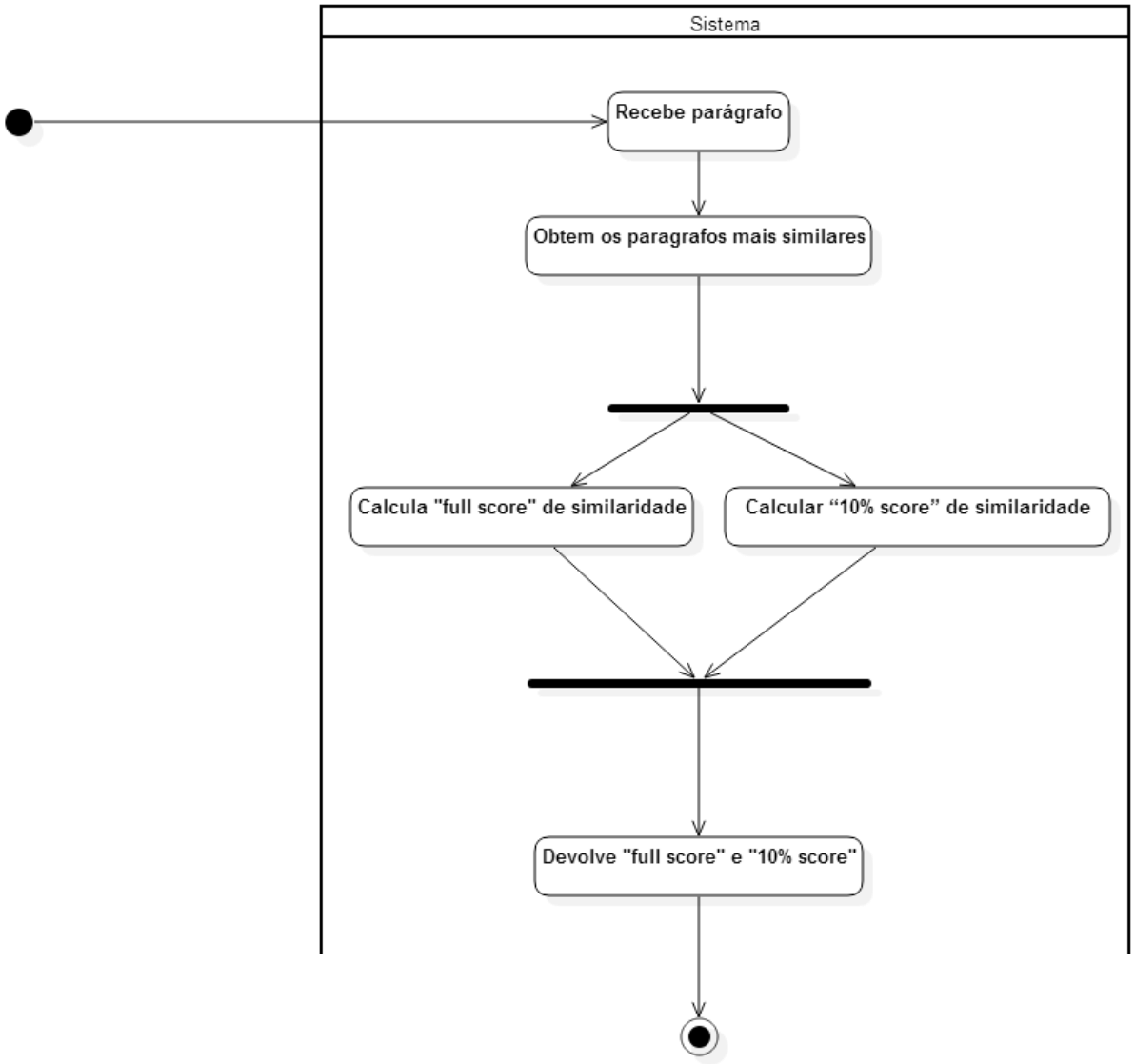


Figura 7.14 – Diagrama de Atividade “Obter Score de Similaridade do Paragrafo”

Caso de Uso: Segmentação de Artigos

Fluxo de Eventos:

1. O sistema recebe Id do artigo
2. Faz caso de uso “Obter Regras de Segmentação”
3. Faz caso de uso “Obter Detalhes do Artigo”
4. Faz caso de uso “Validar Artigo”
5. Faz caso de uso “Obter Lista de Entidades”
6. Classificar artigo
7. O caso de uso termina com sucesso

Cenários Alternativos:

2.a. O caso de uso “Obter Regras de Segmentação” retorna uma lista vazia

2.a.1. O sistema classifica o artigo sem interesse

2.a.2. O caso de uso termina com sucesso

4.a. O caso de uso “Validar Artigo” retorna artigo inválido

4.a.1. O sistema classifica o artigo sem interesse

4.a.2. O caso de uso termina com sucesso

5.a. O caso de uso “Obter lista de entidades” retorna lista de entidades vazia

5.a.1. O sistema continua no 6

6.a. Se o sistema classificar o artigo como irrelevante

6.a.1. Faz o caso de uso “Extração de Parágrafos”

6.a.2 Faz o caso de uso “Obter Score de Similaridade do Paragrafo”

6.a.3 Calcular score do global de similaridade do texto

6.a.4 O sistema classifica artigo

6.a.5 O caso de uso termina com sucesso

6.b. O sistema classifica o artigo como relevante:

6.b.1 O sistema retorna o artigo segmentado

6.a.2 O caso de uso termina com sucesso

6.a.1.a O caso de uso “Extração de Parágrafos” devolve lista de parágrafos vazia:

6.a.1.a.1 O sistema retorna o artigo sem interesse

6.a.1.a.2 O caso de uso termina com sucesso

6.a.4.a O score global é superior ao mínimo:

6.a.4.a.1 O sistema retorna o artigo segmentado

6.a.4.b.2 O caso de uso termina com sucesso

6.a.4.b O score global é inferior ao mínimo:

6.a.4.b.1 O sistema retorna o artigo sem interesse

6.a.4.b.2 O caso de uso termina com sucesso

Diagrama de Atividade:

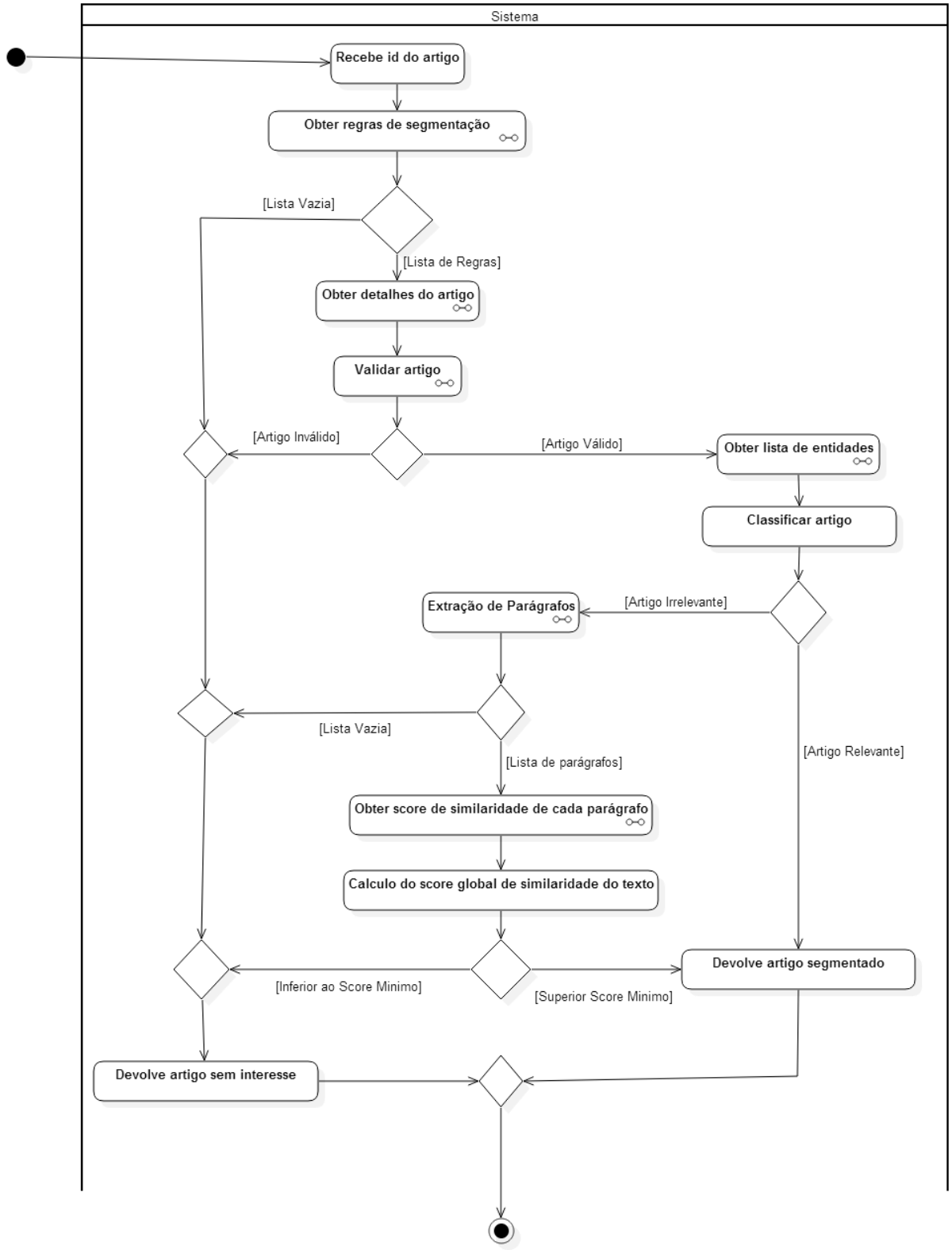


Figura 7.15 – Diagrama de Atividade “Segmentação de Artigos”